Paidós
Básica

Robert Nozick
La naturaleza de la racionalidad

PB

74

La naturaleza de la racionalidad

Paidós Básica

Últimos títulos publicados:

33. G. Duby - *Europa en la Edad Media*
34. C. Lévi-Strauss - *La alfarera celosa*
35. J. W. Vander Zanden - *Manual de psicología social*
36. J. Piaget y otros - *Construcción y validación de las teorías científicas*
37. S. J. Taylor y R. Bogdan - *Introducción a los métodos cualitativos de investigación*
38. H. M. Feinstein - *La formación de William James*
39. H. Gardner - *Arte, mente y cerebro*
40. W. H. Newton-Smith - *La racionalidad de la ciencia*
41. C. Lévi-Strauss - *Antropología estructural*
42. L. Festinger y D. Katz - *Los métodos de investigación en las ciencias sociales*
43. R. Arrillaga Torrens - *La naturaleza del conocer*
44. M. Mead - *Experiencias personales y científicas de una antropóloga*
45. C. Lévi-Strauss - *Tristes trópicos*
46. G. Deleuze - *Lógica del sentido*
47. R. Wuthnow - *Análisis cultural*
48. G. Deleuze - *El pliegue. Leibniz y el barroco*
49. R. Rorty, J. B. Schneewind y Q. Skinner - *La filosofía en la historia*
50. J. Le Goff - *Pensar la historia*
51. J. Le Goff - *El orden de la memoria*
52. S. Toulmin y J. Goodfield - *El descubrimiento del tiempo*
53. P. Bourdieu - *La ontología política de Martin Heidegger*
54. R. Rorty - *Contingencia, ironía y solidaridad*
55. M. Cruz - *Filosofía de la historia*
56. M. Blanchot - *El espacio literario*
57. T. Todorov - *Crítica de la crítica*
58. H. White - *El contenido de la forma*
59. F. Rella - *El silencio y las palabras*
60. T. Todorov - *Las morales de la historia*
61. R. Koselleck - *Futuro pasado*
62. A. Gehlen - *Antropología filosófica*
64. R. Rorty - *Ensayos sobre Heidegger y otros pensadores contemporáneos*
67. A. Schütz - *La construcción significativa del mundo social*
68. G. E. Lenski - *Poder y privilegio*
69. M. Hammersley y P. Atkinson - *Etnografía. Métodos de investigación*
70. C. Solís - *Razones e intereses*
71. H. T. Engelhardt - *Los fundamentos de la bioética*
72. F. Rabossi y otros - *Filosofía de la mente y ciencia cognitiva*
73. J. Derrida - *Dar (el) tiempo. I. La moneda falsa*
74. R. Nozick - *La naturaleza de la racionalidad*
75. B. Morris - *Introducción al estudio antropológico de la religión*
76. D. Dennett - *La conciencia explicada. Una teoría interdisciplinar*
79. R. R. Aramayo, J. Muguerza y A. Valdecantos - *El individuo y la historia*

Robert Nozick

La naturaleza de la racionalidad



ediciones PAIDOS
Barcelona
Buenos Aires
México

Título original: *The nature of rationality*

Publicado en inglés por Princeton University Press, New Jersey

Traducción de Antoni Domènech

Cubierta de Mario Eskenazi

1ª edición, 1995

ISBN: 84-493-0138-6
Depósito legal: B-30.150/1995

Impreso en Hurope, S.L.,
Recaredo, 2 - 08005 Barcelona

Impreso en España - Printed in Spain

Para Carl Hempel
Y A LA MEMORIA DE
Gregory Vlastos

SUMARIO

AGRADECIMIENTOS

Los dos primeros capítulos de este libro fueron originalmente dictados como conferencias Tanner en la Universidad de Princeton, los días 13 y 15 de noviembre de 1991. Yo me gradué en Princeton, así que dediqué esas conferencias, como luego el libro entero, a los que allí fueron mis maestros. Los capítulos 1 y 2 se reproducen con el permiso de la editorial de la Universidad de Utah, y proceden de *Tanner Lectures on Human Values*, vol. 14 (Salt Lake City; University of Utah Press, © 1992). (Las versiones aquí publicadas contienen algunos añadidos y cambios.) Los primeros manuscritos de esos dos capítulos fueron escritos en el centro de investigación de la Fundación Rockefeller en Belagio, Italia, en el verano de 1989.

Partes del capítulo 5 componen la conferencia Walter C. Schnackenberg Memorial, pronunciada en la Universidad Luterana del Pacífico en marzo de 1990. Partes de los capítulos 3-5 se pronunciaron en una conferencia conmemorativa del centenario de la Universidad de Chicago en mayo de 1992.

Estoy en deuda con quienes participaron en las discusiones de las conferencias dictadas en Princeton —Gilbert Harman (que también leyó el manuscrito completo), Clifford Geertz, Susan Hurley y Amos Tversky—, así como con Scott Brewer, Eugene Goodheart, David Gordon, Christine Korsgaard, Elijah Millgram, Bill Puka, Tim Scanlon, Howard Sobel y William Talbott por sus muy útiles comentarios y sugerencias. Particular agradecimiento guardo a Amartya Sen por muchas discusiones estimulantes del presente material, dentro y fuera de las clases que hemos impartido juntos.

Estoy muy reconocido a Laurance Rockefeller por su interés y generosa ayuda a este proyecto de investigación.

Agradezco a mi esposa, Gjertrud Schnackenberg, que convirtiera los años en que se escribió este libro en algo tan romántico y amoroso —y tan divertido—.

INTRODUCCIÓN

La palabra *filosofía* significa amor a la sabiduría, pero lo que los filósofos realmente aman es razonar. Formulan teorías y arman razones para defenderlas, consideran objeciones y tratan de darles respuesta, construyen argumentos contra otras concepciones. Incluso los filósofos que proclaman las limitaciones de la razón —los escépticos griegos, David Hume, quienes ponen en duda la objetividad de la ciencia—, todos aducen razones en favor de sus puntos de vista y presentan dificultades a los puntos de vista opuestos. Las proclamas o los aforismos no se consideran filosofía a menos que entrañen y esbocen razonamientos.

Una de las cosas sobre las que los filósofos razonan es el razonamiento mismo. ¿A qué principios debería obedecer? ¿A qué principios obedece? Aristóteles inició la formulación y el estudio explícitos de los principios deductivos, los autores dedicados a la ciencia y a la teoría de la probabilidad esbozaron modos de razonamiento y apoyo no deductivos, Descartes intentó mostrar por qué deberíamos confiar en los resultados del razonamiento, Hume cuestionó la racionalidad de hacerlo, y Kant acotó lo que tomamos por dominio propio de la razón. Ese bosquejo de la razón no era un ejercicio académico. Los descubrimientos tenían que ser aplicados: el razonamiento de la gente tenía que ser mejorado, sus creencias, sus prácticas y sus acciones tenían que hacerse más racionales. Inquirir en la racionalidad de las creencias y las prácticas coetáneas trae consigo riesgos, como descubrió Sócrates. Las tradiciones de una sociedad a veces no resisten el escrutinio; no todo el mundo desea el examen explícito de lo que anda implícito. Aun la simple consideración de alternativas puede parecer una corrosiva socavación de lo realmente existente, una exposición a la arbitrariedad.

La racionalidad, sostuvieron los griegos, era lo distintivamente humano. «El hombre es un animal racional.» La capacidad para ser racional distingue a los hombres de los otros animales y, así, les define. Lo que haya que considerar específicamente humano ha ido contrayéndose desde la Edad Media. —Ésa es la primera gran afirmación sobre la historia intelectual que yo recuerdo haber leído.— Copérnico, Darwin y Freud nos enseñaron que los seres humanos no ocupan un lugar especial en el universo, no son especiales por

su origen, ni siempre se guían por motivos, no ya racionales, sino meramente conscientes. Lo que continúa dando a la humanidad algún tipo de estatus especial es, a pesar de todo, su capacidad para la racionalidad. Quizá no ejerzamos consistentemente ese valioso atributo; sin embargo, él nos distingue y nos pone aparte. La racionalidad nos dota del poder (potencial) para investigar y descubrir cualquier cosa y todas y cada una de las cosas; nos permite controlar y dirigir nuestra conducta a través de razones y de la utilización de principios.

La racionalidad es, por ende, un componente crucial de la imagen que de sí propia tiene la especie humana, no simplemente una herramienta para ganar conocimiento o mejorar nuestras vidas y nuestra sociedad. Comprender nuestra racionalidad significa inteligir más profundamente nuestra naturaleza y nuestro estatus social —cualquiera que sea—. Los griegos vieron la racionalidad como algo independiente de la animalidad, no desde luego como resultado o culminación de la misma. La teoría evolucionaria permite, en cambio, ver la racionalidad como un rasgo animal entre otros, como una adaptación evolucionaria, con propósito y función bien delimitados.

Esa perspectiva puede traer consigo importantes consecuencias para la filosofía, según creo. La racionalidad no ha sido meramente el amor especial de los filósofos, y una parte importante de su objeto de estudio; también ha sido su herramienta especial para descubrir la verdad, una verdad potencialmente ilimitada. (En la *Crítica de la razón pura,* Kant otorga a la razón una función humilde: no conocer el corazón de una realidad independiente, sino conocer un reino empírico parcialmente constituido y moldeado por ella, por la razón. Con todo, el alcance de su validez seguía siendo extremadamente amplio.) Si la racionalidad es una adaptación evolucionaria con un propósito y una función delimitados, diseñada para colaborar con otros hechos estables que toma como dados y a partir de los cuales construye, pero si la filosofía es un intento, de alcance ilimitado, de aplicación de la razón y de justificación racional de cualquier creencia y de cualquier supuesto, entonces podemos entender por qué muchos de los problemas tradicionales de la filosofía se han revelado indóciles y resistentes a su resolución racional. Quizá esos problemas son el resultado de la pretensión de extender la racionalidad más allá de su función, evolucionariamente delimitada. Me refiero aquí a problemas como el de la inducción, el de las otras mentes, el del mundo externo y el de la justificación de los fines. Más adelante exploraré las consecuencias y las implicaciones de esa perspectiva evolucionaria.

En los últimos años, la racionalidad se ha convertido en un particular objeto blanco de las críticas. Se ha adelantado la idea de que la racionalidad está *sesgada* porque es una noción de clase, o masculina, u occidental, o cualquier cosa por el estilo. Sin embargo, es parte de la racionalidad el advertir sesgos, incluidos los suyos propios, y controlarlos y corregirlos. (¿Podría ser el intento mismo de corregir sesgos un sesgo más? Mas, si eso fuera una *crítica,* ¿de dónde procede esa crítica? ¿Se trata de un punto de vista que afirma que los sesgos son malos, pero que corregirlos es malo también? Si se sostiene que es imposible eliminar los sesgos, entonces ¿en qué sentido imputar sesgos constituye una crítica? ¿Y significaría tal imposibilidad que hay algún tipo particular de sesgo que es intrínsecamente resistente a la eliminación, o sólo que no todos los sesgos pueden eliminarse simultáneamente?)

Imputar un sesgo a los criterios existentes no demuestra que exista tal sesgo. Para demostrarlo hay que usar el razonamiento y la evidencia —haciendo, por tanto, uso de los criterios existentes para llegar a la conclusión de que esos criterios mismos, en algunas de sus aplicaciones, revelan algunas distorsiones o algunos sesgos específicos particulares—. No basta con limitarse a decir que (todos) nosotros vemos el mundo a través de nuestros esquemas conceptuales. La cuestión es: ¿de qué modos concretos y a través de exactamente qué mecanismos llevan nuestros particulares esquemas conceptuales y criterios a la distorsión? Una vez se nos ha mostrado eso, podemos empezar a introducir correcciones. Ni que decir tiene que nuestros criterios habituales de racionalidad no son perfectos (¿en qué año se supone que adquirieron perfección?). Mas poseen auténticas virtudes, y para probar que no están sin mácula es necesaria la argumentación racional, o al menos, aportar un peso parecido al de los criterios atacados. Detectar máculas particulares de este tipo es el primer paso necesario para repararlas y para formular de un modo más adecuado los criterios de racionalidad. De manera que habría que dar la bienvenida a, y buscar activamente, la evidencia probatoria que permite imputar sesgos a los criterios. Los criterios de racionalidad son un medio por el que superamos o moderamos nuestros propios deseos, esperanzas y sesgos particulares. Sería a la vez irónico y trágico que la presentemente extendida suspicacia crítica frente a los criterios de racionalidad tuviera el efecto de liquidar o socavar una de las vías capitales por las que a la humanidad le es dado corregir y superar sesgos personales y colectivos.

El estudio de la racionalidad, de tanta importancia evaluativa y práctica —tanto personal, cuanto socialmente—, se ha transforma-

do en un asunto técnico. Los principios se han ido afilando para perfilar el razonamiento válido y capturar conceptualmente los patrones de conducta y de acción apoyados en razones. La lógica deductiva fue transformada por Gottlob Frege a finales del siglo diecinueve y ha experimentado un estallido de elaboración técnica en el siglo veinte. Se han desarrollado sistemas de lógica y se han explorado sus propiedades y limitaciones usando técnicas lógicas. La teoría de la probabilidad ha llevado a teorías formales de la inferencia estadística, y la matematización ha permeado los intentos de teorizar acerca de la racionalidad de la creencia y de formular los rudimentos de una lógica inductiva, o al menos, reglas inductivas de aceptación. Una teoría bruñida y potente de la acción racional —la teoría de la decisión— ha sido desarrollada a lo largo de este siglo por matemáticos, economistas, estadísticos y filósofos, y ahora aplicamos esa teoría a una amplia variedad de contextos teóricos y prácticos. (El aparato de esa teoría suministra el marco conceptual para la teoría formal de la interacción estratégica racional, la teoría de los juegos, la teoría formal de la elección social y la economía de bienestar, la teoría de los fenómenos microeconómicos y elaboradas teorías del dominio de la política.) La literatura relevante está salpicada por, sino enteramente sumergida en, prohibitivas fórmulas escritas en raras notaciones simbólicas con las que se elaboran estructuras matemáticas. No lamento ese giro. Estos desarrollos actuales están en solución de continuidad con las tempranas motivaciones y preocupaciones, y constituyen un progreso muy importante en la investigación.

También este libro tomará en cuenta tales tecnicismos y propondrá algunos nuevos en las dos áreas principales cubiertas por las teorías de la racionalidad: la racionalidad de la decisión y la racionalidad de la creencia. Reformaremos la teoría actual de la decisión para incluir en ella el significado simbólico de las acciones, propondremos una nueva regla de la decisión racional (la de maximizar el valor decisional) y buscaremos las implicaciones de esa regla para el dilema del prisionero y para el problema de Newcomb. La racionalidad de la creencia entraña dos aspectos: apoyo en razones que hacen creíble a la creencia, y generación mediante un proceso seguro de producción de creencias verdaderas. (La idea evolucionaria que ofrezco para explicar la enigmática conexión entre esos aspectos invierte la dirección de la «revolución copernicana» de Kant.) Propondré dos reglas para gobernar la creencia racional: no creer ningún enunciado menos creíble que alguna alternativa incompatible —el componente intelectual—, pero creer un enunciado sólo si la utili-

dad esperada (o el valor de la decisión) de hacerlo es mayor que el de no creer en él —el componente práctico—. Esa estructura doble se aplica a asuntos relacionados con la «ética de la creencia», y se propone una nueva solución a la «paradoja de la lotería». También exploraré el alcance y los límites de la racionalidad instrumental, la efectiva y eficiente persecución de fines dados, y propondré algunas condiciones nuevas para la racionalidad de los fines. Puesto que el pensamiento racional comprende también la formulación de cuestiones e ideas filosóficas nuevas y fértiles, se ofrecerá asimismo cierta heurística pertinente para ello. De manera que este libro rebosa en detalles técnicos necesarios para hacer avanzar el pensamiento sobre asuntos de racionalidad.

Mas no faltan motivos de preocupación al respecto. Hasta hace poco, las cuestiones acerca de la racionalidad habían sido propiedad común de la humanidad, a veces discutidas por intrincadas veredas del pensamiento —nadie puede decir que la *Crítica de la razón pura* sea un libro fácil—, pero, sin embargo, holgadamente accesibles a un público inteligente dispuesto a hacer el esfuerzo. Ideas nuevas sobre esas cuestiones ingresaban pronto en la cultura general; modelaban los términos de la discusión y del debate, y a veces incluso la sensibilidad (recuérdese hasta qué punto influyó en Coleridge el pensamiento de Kant). Ahora las cosas son diferentes, y no sólo respecto del tópico de la racionalidad.

Las líneas más fértiles e interesantes de investigación acerca de varios asuntos de interés humano fundamental han experimentado un giro crecientemente técnico. Es imposible discutir ahora adecuadamente esos asuntos sin tener alguna idea de esos desarrollos técnicos, de las nuevas cuestiones abiertas por ellos, y de la forma en que las posiciones tradicionales han sido socavadas. Cuando la Enciclopedia Británica publicó recientemente su (segunda) edición de *Los grandes libros del mundo occidental,* dio pie a cierta controversia sobre la representación —o relativa carencia— de mujeres y minorías, y sobre el supuesto elitismo de cualquier canon de grandes obras.* Lo que, sin embargo, no mereció comentario alguno es que muchas de las grandes obras intelectuales del siglo veinte fueran

* Por mi parte, no me parece que una edición uniforme de las obras de muchos autores diferentes, con los títulos de colección más prominentemente destacados que los títulos de las obras individuales o los nombres de los autores, sea una presentación adecuada de los logros escritos conseguidos por la mente. Podría ser útil, no obstante, que un grupo publicara una *lista* de tales libros y reimprimiera los que no fueran fácilmente accesibles; distintos grupos podrían publicar distintas listas.

omitidas, presumiblemente porque eran demasiado técnicas para el lector inteligente educado en una cultura general.

No se trata sólo de que los pensamientos y los resultados interesantes acontecidos en el presente siglo sean inaccesibles para amplios sectores de una población instruida —eso ha sido así desde Newton—. Ocurre, más bien, que ahora esas ideas tienen que ver con asuntos que deseamos y necesitamos entender, con asuntos de los que pensamos que deberían poder ser entendidos por todo el mundo. Sin embargo, sin alguna familiaridad con los tecnicismos, esos asuntos no pueden ser comprendidos ni discutidos inteligentemente. Los términos mismos de la evaluación se han hecho técnicos.

Permítaseme dar algunos ejemplos de asuntos que han experimentado un desarrollo técnico. (1) La noción de bienestar general (y la noción rousseauniana de «voluntad general») y una comprensión de los propósitos de los procedimientos de votación democrática han sido transformadas por el teorema de imposibilidad de Kenneth Arrow. Muestra éste que varias condiciones extremadamente naturales y deseables, que aparentemente deberían ser satisfechas por cualquier procedimiento tendente a determinar el bienestar general o la alternativa democráticamente preferida, no pueden satisfacerse simultáneamente. Alguno debe ser abandonado. (2) El trabajo de Amartya Sen sobre la paradoja del paretiano liberal muestra que una interpretación muy natural del alcance de los derechos y libertades individuales no casa fácilmente con el modo en que habría que organizar racionalmente las elecciones de la sociedad. Esas nociones necesitan una nueva estructuración. (3) La estructura fundamental del mundo físico —la estructura del espacio y del tiempo— no resulta ya inteligible si se prescinde de los tecnicismos (y las matemáticas) del espacio-tiempo presentes en la teoría general de la relatividad. (4) Algo análogo acontece con la naturaleza de la causalidad y con el carácter independiente del mundo físico, tal como los perfila la teoría científica más precisa y exitosa de que disponemos hoy, la teoría cuántica de campos. (5) La discusión sobre la naturaleza y el estatus de la verdad matemática —desde los griegos, el paradigma de nuestro mejor y más cierto conocimiento— ha sido drásticamente transformada por los teoremas de incompletud de Kurt Gödel. (6) La naturaleza del infinito y de sus varios niveles se elabora y explora hoy en la teoría contemporánea de conjuntos. (7) Sin la teoría que estudia la manera en que un mecanismo de precios, y las instituciones de propiedad privada que le acompañan, hacen posible el cálculo económico racional, sin la discusión téorica —que duró décadas— acerca de si era posible el cálculo racional en una

sociedad socialista, no puede entenderse por qué las sociedades comunistas fueron económicamente tan ineficaces. (8) En lo que hace a la racionalidad individual y a las interacciones racionales entre las personas, ha habido muchos progresos teóricos: la teoría de la decisión, la teoría de los juegos, la teoría de la probabilidad y las teorías de la inferencia estadística.

En cada uno de los asuntos mencionados, este siglo ha visto resultados y teorías espectacularmente nuevos, resultados y teorías que son difíciles de comprender o de discutir responsablemente sin una inteligencia de las estructuras y los detalles técnicos. Se trata, no se me escapa, de la lista de un filósofo; los científicos sociales y naturales añadirían otros asuntos. Lo que no hace sino dar mayor pábulo a lo que estoy sosteniendo. La cultura común de las personas inteligentes, instruidas y serias ha perdido pie en muchos asuntos que resultan centrales para inteligir o pensar acerca de la sociedad, o la gente, o el universo entero. La tesis de que hay factualidades científicas problemáticamente complicadas cuya resolución hay que encargar a los expertos, expertos que, a su vez, estarán en desacuerdo (hechos, por ejemplo, acerca de los efectos medioambientales de varias prácticas), resulta ya familiar. Lo nuevo es esto: muchos de los términos y conceptos que *nosotros* queremos usar para evaluar y comprender se han convertido, ellos también, en términos y conceptos técnicos.

Levanto esta cuestión sin tener una solución que proponer. Obvio es decir que el público general necesita exposiciones de esos asuntos. Pero ni la más clara de ellas, si quiere transmitir las ideas esenciales con cierto esmero, podrá evitar algunas descripciones y algunos desarrollos técnicos —lo que limitará su público—. La tarea es aún más difícil para una obra que presenta y explora ideas nuevas. Yo no *deseo* que el asunto de la racionalidad se hurte al público general. Pero algunas ideas sólo pueden formularse, definirse o defenderse de una manera más o menos técnica. He tratado de minimizar esos detalles técnicos, o confinarlos al menos a secciones específicas. Para la salud intelectual de nuestra sociedad —y no digamos para la salud social de nuestros intelectuales— es imprescindible que las ideas fundamentales sigan siendo públicas.

CAPÍTULO 1

CÓMO HACER COSAS CON PRINCIPIOS

¿Para qué sirven los principios? ¿Por qué sostenemos principios, por qué los proponemos, por qué nos adherimos a ellos? Podríamos simplemente actuar antojadizamente, o según la pasión del momento, o podríamos limitarnos a maximizar nuestro propio interés y recomendar a los otros lo mismo. ¿Son acaso los principios una restricción al antojo y al interés egoísta? ¿O es más bien la adhesión a principios un modo de promover el interés propio? ¿A qué funciones sirven los principios?

Los principios de acción agrupan las acciones, poniéndolas bajo rótulos generales; así, las acciones conectadas se ven o se tratan de la misma manera. Esa generalización puede servir a diferentes funciones: intelectuales, interpersonales, intrapersonales y personales. Empezaré con las intelectuales.

LAS FUNCIONES INTELECTUALES

Consideremos la toma de decisiones judiciales. En un sistema imaginable, un juez simplemente decide un caso con objeto de conseguir lo que cree que es el resultado mejor o preferible en un caso particular. Otro sistema de decisión judicial entraña una decisión de principios: un juez de derecho consuetudinario angloamericano tiene que formular un principio que case con (la mayoría, o casi todos) los precedentes y con una gama de casos hipotéticos, y luego usar ese principio para decidir el caso que tiene entre manos.* El

* Me propongo aquí iluminar algunos rasgos generales que tienen los principios fuera del ámbito jurídico por analogía con algunos aspectos de la decisión judicial, no presentar una imagen completa del funcionamiento de las instituciones jurídicas. Lo que resulta ilustrativo es la analogía entre el modo en que una decisión judicial corriente tiene que salir de un principio que case con los precedentes y el modo en que (fuera del derecho) un principio tiene que arrojar juicios correctos. Que dentro del sistema jurídico *stare decisis* sea por sí mismo un principio (de orden superior) del derecho, un principio que puede a veces entrar en conflicto o competir con otros principios, es cosa que carece ahora de interés para nosotros.

intento de formular un principio general aceptable constituye un *test* para el juicio que hagan ustedes acerca del caso particular: ¿hay *algún* principio general adecuado —un principio que arroje el resultado correcto en todos los casos pasados y en los casos hipotéticos obvios— que también arroje el resultado que ustedes quieren en el caso en cuestión? Si ustedes no pueden hallar tal principio, reconsideren el resultado que desean en este caso.

Tal procedimiento constituye un test para un juicio particular suponiendo que cualquier juicio correcto salga de *algún* principio general aceptable, que los juicios particulares verdaderos sean consecuencias de principios generales aplicados a situaciones concretas. Un fracaso a la hora de formular un principio general aceptable que arroje algún juicio en particular puede significar que no existe tal principio aceptable, en cuyo caso tal juicio particular anda errado y debería ser abandonado. O quizá no han sido ustedes lo suficientemente astutos como para formular el principio correcto. No disponemos de ningún procedimiento mecánico para decidir qué explicación es la correcta.[1]

Cuando ustedes hallan un principio o una teoría general que subsume este caso, un principio que ustedes estuvieran dispuestos a aplicar a otros casos también, este juicio particular recibe apoyo nuevo. Considérense los datos empíricos constituidos por los puntos, *a, b, c, d.* Si una línea recta es la curva más sencilla que pasa a través de esos puntos, eso da apoyo a la predicción de que otro punto *e*, también en la línea recta, ocurrirá. Los lógicos inductivos han descubierto que no es cosa fácil aislar y explicar el modo en que un enunciado legaliforme (relativamente) simple puede agrupar datos de puntos de manera tal que puedan legítimamente inferirse y hacerse predicciones sobre nuevos puntos. No obstante, no dudamos que los datos pueden apoyar la hipótesis de que una ley está vigente y la predicción de que habrá un nuevo punto acorde con la ley. Análogamente, el principio más simple que cubre los puntos normativamente aceptables *a, b, c, d,* confiere también apoyo a un juicio adicional sobre *e* (que casa con ese principio) como un punto normativamente correcto también. Un teórico gana confianza en su juicio particular (o en el partido que toma en una controversia) cuando puede formular un principio o teoría general que case con él, sobre todo si su misma apariencia resulta atractiva.[2]

Los filósofos de la ciencia han intentado trazar una línea de demarcación entre las leyes científicas y las generalizaciones accidentales. A las generalizaciones accidentales sólo les acontece que son, o han sido, verdaderas. De una generalización como, por ejemplo,

todas las monedas que están en mi bolsillo son de diez centavos, no se puede inferir un enunciado subjuntivo como el siguiente: Si *hubiera* otra moneda en mi bolsillo ahora, *sería* una moneda de diez centavos. (De una ley científica, en cambio —verbigracia: que todos los cuerpos en caída libre caen a una distancia igual a $1/2gt^2$—, podemos inferir que si algún objeto ahora en reposo cayera libremente durante t segundos, viajaría a una distancia igual a $1/2gt^2$.) Si todos los datos previos casan con una generalización dada, podemos inferir plausiblemente que los nuevos datos casarán también con ella (y por lo tanto, predecir que los nuevos datos que serán colectados casarán también, *sólo* si esa generalización es legaliforme y es candidata a ser una ley. Sólo cuando los datos caen bajo un enunciado legaliforme (o surgen de varios de ellos) podemos legítimamente extrapolar a casos ulteriores. Los rasgos de un enunciado legaliforme, los aspectos que lo distinguen de una generalización accidental, nos autorizan a viajar desde los datos hasta las predicciones o expectativas de datos ulteriores. Análogamente, en el caso de los juicios normativos particulares, lo que nos autoriza a viajar hasta un juicio ulterior partiendo de juicios previos es el hecho de que éstos caigan bajo un juicio normativo general. Los rasgos de un juicio normativo autorizan una inferencia subjuntiva hasta un nuevo caso que va más allá de los casos indicativos que ya habían caído bajo ese juicio. Los principios son mecanismos de transmisión de *probabilidad* o *apoyo,* los cuales van desde los datos o los casos, a través de los principios, hasta juicios y predicciones acerca de nuevas observaciones o casos cuyo estatus resulta, por otra parte, desconocido o menos cierto.

¿Qué rasgos permiten a los principios transmitir probabilidad? Los siguientes rasgos han sido mencionados para distinguir enunciados científicos legaliformes (o universales nómicos) de generalizaciones accidentales.[3] Los enunciados legaliformes no contienen términos para objetos, fechas o períodos temporales individuales particulares —o, si los contienen, pueden derivarse de enunciados legaliformes más generales que no los contienen—. Los enunciados legaliformes contienen predicados puramente cualitativos: fijar el significado de éstos no implica referencia a ningún objeto o localización espacio-temporal *particular.* Los enunciados legaliformes poseen una universalidad irrestricta; no son simplemente una conjunción finita conseguida tras examinar todos los casos. Los enunciados legaliformes reciben su apoyo no de casos que caen bajo ellos, sino de una concatenación de evidencia indirecta.

Son esos mismos rasgos los que permiten que un principio nor-

mativo autorice una derivación de juicios nuevos a partir de los previamente aceptados. Los especialistas en ética suelen decir que los principios éticos deben formularse usando exclusivamente términos generales —nada de nombres de personas, grupos o naciones particulares—. Ese rasgo puede permitir a un principio autorizar una inferencia de un caso nuevo, permitiendo así que los juicios normativos nuevos reciban apoyo de los previos. Una generalización que carezca de ese rasgo de no-particularidad podría, a lo sumo, ser una generalización accidental, incapaz de transferir apoyo de unos datos a otros. El que los principios morales sean generales y no contengan predicados no-cualitativos o nombres particulares de tipo alguno, más que ser un aspecto específicamente *moral* de los principios, podría ser un rasgo requerido para concatenar datos o juicios con objeto de dar apoyo a inferencias subjuntivas. Valdría la pena investigar hasta qué punto la «forma» de los principios morales resulta necesaria para tal concatenación.

Eso no significa que esos rasgos se añadan a generalizaciones más débiles para conseguir que los principios morales realicen funciones inferenciales; tampoco se añaden esos rasgos a las generalizaciones accidentales para convertirlas en leyes científicas. Se puede sostener que las leyes científicas y los principios morales se mantienen verdaderos independientemente de cualesquiera construcciones que les añadamos o de cualquier uso que de ellos hagamos, que la independencia de su verdad es lo que hace posibles esos usos. No obstante, rasgos como los de generalidad, ausencia de nombres propios y de predicados posicionales no serían rasgos específicamente *morales*, sino legaliformes, necesarios para que algo sea una ley, científica o moral. En el contexto apropiado, rasgos que no son específicamente morales pueden tener consecuencias morales.

Una persona puede andar a la búsqueda de principios no sólo para someter a test su propio juicio o para conferirle mayor apoyo, sino para convencer a otros, o para robustecer su propia convicción. No le bastará para ello con anunciar su preferencia por una posición; deberá producir razones que resulten convincentes para otros. Las razones pueden ser de carácter muy particular, pero pueden ser también consideraciones generales aplicables a una amplia gama de casos, interesando en esta ocasión a un juicio particular. Si esos juicios resultan aceptables para otros en otros casos, entonces el razonamiento general reclutará esos casos como evidencia y apoyo para el juicio que propone en el caso presente. Los principios o las teorías generales tienen, pues, una función intelectual general: la justificación ante otros. La justificación mediante principios generales

resulta convincente de dos maneras: por el atractivo aparente de los principios y por reclutar otros casos ya aceptados en apoyo de una posición propuesta para el caso presente.[4]

Al usar un juez a modo de ilustración de la función de someter a test y apoyar que tienen los principios, he imaginado que su propósito es el de llegar a la decisión correcta en un caso particular y que cree (en su mayor parte) correctas las decisiones pasadas. Es decir, he considerado al juez como estructuralmente idéntico al razonador moral que quiere decidir lo que es correcto o permisible en esta nueva ocasión o situación y que usa su conocimiento de lo que es correcto o permisible en otras situaciones, reales o hipotéticas, para formular, someter a test y apoyar un principio moral que arroja un resultado para esta situación.

Claro es que un juez es también una figura en una estructura institucional, y las decisiones de principio que casan con los casos pasados pueden tener una particular justificación dentro de esa institución. Los teóricos del derecho nos enseñan que la doctrina del respeto de los precedentes, *stare decisis*, permite a la gente predecir más exactamente las decisiones futuras del sistema jurídico y planear así las acciones con cierta seguridad respecto de sus consecuencias jurídicas.[5] A este efecto, no es necesario que los precedentes hayan sido decididos correctamente o sean seguidos con el fin de llegar a una decisión justa; son seguidos para conseguir el resultado esperado. En segundo lugar, las decisiones que apelan a principios pueden ser deseables porque restringen la base decisional del juez. Se trata de excluir sus preferencias o prejuicios personales, sus humores momentáneos, su parcialidad en una disputa, o incluso los principios morales y políticos arraigados y personales. Podría sostenerse que los puntos de vista, las preferencias y hasta las convicciones más mediatas de un juez no deberían tener más efecto que los de cualquier otra persona —al juez no se le ha otorgado su posición institucional para que sus preferencias tengan efectos—. La exigencia de que las decisiones actuales casen mediante principios con las precedentes podría ser un mecanismo para *restringir* el efecto de tales factores personales, limitando su juego o eliminándolo *por completo.*

Sin embargo, la analogía con la ciencia, donde el propósito es la verdad y la exactitud, arroja dudas sobre esta última y estricta exigencia. Casar con los datos científicos es una exigencia, pero eso no determina de un modo unívoco un enunciado legaliforme (incluso dejando de lado el asunto de que hay un amplio margen para definir lo que sea «el mejor modo de casar»). Un número indefinido

de curvas puede casar con cualquier conjunto finito de datos de puntos; más de una sería legaliforme. De manera que serán necesarios criterios adicionales para seleccionar qué enunciado legaliforme hay que aceptar tentativamente para usarlo en la predicción. Esos criterios incluyen la simplicidad, la analogía con enunciados legaliformes que gocen ya de apoyo en áreas vecinas,[6] el casar con otras teorías aceptables, el poder explicativo, la fertilidad teórica y quizá la facilidad de computación.[7] Limitarse a exigir que una predicción case con los datos del pasado de acuerdo con algún enunciado legaliforme no determina unívocamente tal predicción. ¿Hasta qué punto es entonces verosímil que el mero requisito de que la decisión de un juez case con las decisiones pasadas de acuerdo con algún principio basta para determinar esa decisión unívocamente? En realidad, hay jueces a los que se les prescribe usar criterios adicionales, incluidos algunos criterios «formales».[8] También pueden plantearse cuestiones parecidas en el ámbito de la ética. W.V. Quine sostiene que la totalidad de los (posibles) datos empíricos no determina unívocamente una teoría explicativa. ¿Están los principios éticos correctos unívocamente determinados por la totalidad de los juicios correctos acerca de los casos particulares, reales o hipotéticos, o impera también aquí la subdeterminación? Además de casar con juicios particulares, ¿debe un principio moral satisfacer también ciertos criterios adicionales?

Hay una conexión entre usar principios como mecanismos para alcanzar decisiones correctas y usarlos para restringir la influencia de factores indeseados o irrelevantes como las preferencias personales. Queremos decidir o juzgar un caso particular considerando todas las razones relevantes —sólo las relevantes— pertinentes. Un principio general que nos obliga a atender a otros casos, reales o hipotéticos, puede ayudarnos a comprobar si una razón *R* que creemos relevante o concluyente en este caso lo es realmente. ¿Sería *R* relevante o concluyente en otro caso? Si las razones son generales, podemos comprobar la fuerza de *R* en este caso considerando otros casos. Además, decidir mediante un principio general puede llamar nuestra atención sobre otras razones relevantes de las que aún no nos hayamos percatado en este caso. Atender a otro caso en el que la razón *R no* tiene mucha fuerza podría llevarnos a percibir otro rasgo *F* que caracteriza al caso presente, y a percatarnos de que lo que tiene fuerza es la conjunción de *R* y *F*. (Si no hubiéramos atendido al otro caso, podríamos haber pensado que bastaba con *R*.) Atender a todas las razones relevantes puede contribuir a asegurar que *sólo* se hace uso de las razones relevantes *si* éstas cubren todo el es-

pacio y expulsan a las razones irrelevantes. ¿Y estaremos de verdad dispuestos a aceptar el impacto que una razón irrelevante impuesta en este caso podría también tener sobre otros casos y ejemplos? Obsérvese que este uso de los casos hipotéticos o de otros casos reales para someter a test un juicio en este caso presume ya que las razones son *generales*. Si presumimos que las cosas ocurren o vigen por una razón (o causa) y que las razones (o causas) son generales, entonces puede formularse un principio general, acaso rebatible, para captar esa razón, para explicar por qué acaece un acontecimiento estudiado por un científico, o por qué es correcto un juicio particular acerca de un caso.[9]

Los principios pueden guiarnos hacia una decisión o hacia un juicio correctos en un caso particular, ayudándonos a comprobar nuestro juicio y a controlar los factores personales que podrían descarriarnos. La incorrección de la que deben guarnecernos los principios es, desde este punto de vista, individualista —el juicio incorrecto en *este* caso—, o agregativa —los juicios incorrectos en *esos* casos, que son incorrectos uno a uno—. Sin embargo, tomados conjuntamente, los juicios podrían contener una incorrección adicional, una incorrección *comparativa* que acontece cuando los casos que deberían ser decididos del mismo modo son diferentemente decididos. Se ha sostenido que es una máxima de la justicia (formal) la de que casos parecidos deben decidirse parecidamente; esa máxima general deja abierta la cuestión de qué parecidos son los parecidos relevantes.[10] Los principios podrían funcionar para evitar esa injusticia o disparidad, no simplemente por mor de decidir correctamente todos y cada uno de los casos, sino para decidir similarmente casos relevantemente similares. Pero si yo veo películas dos semanas seguidas, no tengo por qué decidir cuáles ir a ver sobre una base similar. Esas dos decisiones similares, pues, aparentemente no cuentan como casos parecidos que deben decidirse parecidamente. (La primera decisión puede afectar a la segunda elección, pero no la restringe.) ¿Cuál es la frontera de demarcación del dominio en el que ha de operar la máxima de la justicia formal? Como espectador de cine no entiendo mi tarea de decidir qué película ir a ver (en cualquiera de las ocasiones) como una tarea de lograr una decisión *justa* en esta ocasión. El asunto de la injusticia comparativa surge sólo en contextos que entrañan justicia o injusticia individual, independientemente del modo en que se delimiten esos contextos. Si el caso A, que necesita una decisión justa, se decide incorrectamente, eso es un mal. Si ahora el caso B, relevantemente similar, se decide de modo diferente —es decir, correctamente— y *si* esa deci-

sión introduce un mal adicional en el mundo —no el resultado en el caso B mismo, sino el mal comparativo de los dos casos decididos de modo diferente— y ese mal rebasa el mal en que se incurrió al decidir el caso A incorrectamente, entonces *este* contexto de justicia es un contexto comparativo que apela a la máxima formal de justicia.* Una función de los principios, pues, puede ser la de evitar este tipo particular de injusticia, asegurando que casos parecidos sean decididos parecidamente. (Si sería mejor decidir ambos casos incorrectamente —evitando la injusticia comparativa— o decidir uno de ellos correctamente —evitando la injusticia en este caso individual, pero incurriendo en la injusticia comparativa— dependerá presumiblemente de rasgos particulares de la situación y de los casos.)

Las funciones interpersonales

Podemos contar con que una persona dotada de principios se atenga a sus principios aun inducida y tentada a desviarse de ellos. No necesariamente frente a cualquier posible tentación o a una inducción extremadamente grande —pero los principios constituyen una suerte de barrera que estorba a la persecución de los deseos o intereses personales del momento—. Los principios de acción de una persona tienen, pues, una función interpersonal, dado que aseguran a otros que (normalmente) esa persona dejará pasar las tentaciones; también tienen una función intrapersonal, ayudando a la persona misma a vencer la tentación.

* He dicho que una condición necesaria para apelar a la máxima formal de justicia es que el contexto sea un contexto en el que haya que alcanzar una decisión justa, pero no he sostenido que se trate de una condición suficiente. Si hay decisiones individuales que entrañan justicia y carecen de aspecto comparativo, entonces se necesita un criterio adicional para indicar qué contextos que entrañan justicia apelan a la máxima formal. En *Anarchy, State, and Utopia* (Nueva York: Basic Books, 1974), cap. 7, presenté una teoría de la justicia distributiva, la teoría de las titulaciones, que explícitamente renunciaba a ser una teoría pautada o configurada y a entrañar comparaciones entre las posesiones de personas distintas. Eso no quiere decir, sin embargo, que la máxima formal no se aplicara a las posesiones cuyos orígenes estuvieran de acuerdo con los *mismos* principios generales (de la justicia en la adquisición, en la transferencia y en la rectificación). De aquí que, hasta donde llega esta teoría, además de la injusticia de que el origen de las posesiones de alguien esté fuera de la operación de esos principios, podría haber una injusticia comparativa adicional si el origen de las posesiones de otro no estuviera fuera de esas operaciones (por ejemplo, el primero es discriminado frente a otros que no dejan que aquellos principios de justicia en las posesiones se apliquen a él).

Empecemos con la función interpersonal. Cuando los principios de una persona mandan (abstenerse de) una acción, podemos confiar más en ello. Siendo capaces de confiar en algún grado en su conducta, nosotros mismos podemos ejecutar acciones cuyo buen resultado depende de la específica conducta de la persona dotada de principios. Aun si el futuro le deparara incentivos para desviarse, podemos confiar en que no lo hará, y, fundándonos en ello, planificar y ejecutar nuestras propias acciones. De otro modo, tendríamos que actuar de manera diferente, pues la probabilidad de que su conducta previa acabe malográndose o frustrándose sería demasiado grande. Respecto de la gente que nos es suficientemente cercana, podemos confiar en su afecto y en la continuidad de sus buenas motivaciones para producir acciones coordinadas; en lo que hace a otros que nos son más lejanos, confiamos en su conducta orientada por principios.

Tales consideraciones son corrientes en las discusiones del derecho contractual. Los contratos permiten a una persona atarse a sí misma para ejecutar una acción, estimulando así a otra a contar con ello y a ejecutar una acción que la deja suspensa en una rama que sería segada si la primera persona no ejecutara la suya. Puesto que la primera persona se beneficia de la acción de la segunda, que no sería ejecutada si la primera persona no se hubiera atado contractualmente a sí misma para ejecutarla, esa primera persona está dispuesta de antemano a constreñirse a sí misma a actuar de ese modo aun en el caso de que sus incentivos futuros fueran a cambiar. Pues si dejara su actuación a merced de las fluctuaciones futuras, la segunda persona no ejecutaría esa acción complementaria que ella, la primera, desea ahora que se ejecute.

Los principios constituyen una forma de atadura: nos atamos a nosotros mismos para actuar según mandan los principios. Otros pueden depender de esa conducta, y nosotros podemos beneficiarnos también de esa dependencia de otros, pues las acciones que de ese modo ellos están dispuestos a ejecutar pueden facilitarnos las interacciones y el trámite social, así como nuestros proyectos personales.[11] *Anunciar* los principios es un modo de incurrir en (lo que los economistas llaman) efectos de reputación, es hacer de tal modo explícitas las condiciones, que las desviaciones son más fácilmente detectables. Esos efectos son importantes para quien haga repetidas transacciones con mucha gente; se garantiza a los demás que se actuará de cierto modo (con objeto de evitar la mengua de una reputación que es útil en la interacción).[12]

Consideraciones de este tipo son las que pueden hacer que una

persona desee *aparentar* frente a otros que tienen principios particulares, mas ¿por qué habría de desear realmente tenerlos? Para la mayoría de nosotros, tener principios puede ser el camino más convincente y menos difícil de aparentar tenerlos, pero tanto la ficción como la vida real rebosan de peritos en el arte de mentir. Supongamos que una persona desea tener un principio particular, y no meramente aparentar tenerlo, porque eso es lo que resultará más convincente para otros y será lo más fácil para él mismo. ¿*Puede* llegar a adquirir ese principio sólo porque cumpla funciones interpersonales útiles? ¿No debería creer que el principio es *correcto*? (Por lo tanto, ¿no desempeña la función intelectual un papel en la función interpersonal?)

¿Y qué garantía me ofrece a mí alguien que me cuenta que cree que tener un principio es necesario para darme garantías a mí y a otros? «Pero ¿lo *tiene* usted?» preguntaría yo. «¿Y con qué fuerza lo alberga?» Si él concibe el principio simplemente como una garantía para otros, aun una garantía muy necesaria y extremadamente útil, ¿no me maravillaría a mí su indesmayable adhesión a la vista de tentaciones e incentivos momentáneos para desviarse de él? Creo que lo que yo desearía es que la persona creyera que el principio es *correcto* y *justo*. Evidentemente no basta con que ella lo crea así ahora; su creencia debe ser estable, no sujeta a cambio al menor contraargumento o contraincentivo. Eso es lo que me daría a mí garantías suficientes para emprender acciones arriesgadas cuyo buen resultado dependiera de su buena conducta. Y yo podría ser muy eficiente en punto a detectar la genuinidad de una creencia en un principio, y no estar dispuesto a correr riesgos cooperativos en ausencia de ella.[13]

Creer en la corrección de sus principios, pues, podría ser un rasgo útil para una persona, posibilitándole una gama más amplia de interacciones con otros y actividades operativas. Esa creencia podría resultar útil aun si la noción de «principios correctos» no tuviera ningún sentido. Pues esa —supongámoslo por un momento— creencia sin sentido, experimentada por él y detectada por otros, sería un indicador fiable para éstos de la conducta futura de aquél, y les llevaría a acciones de las que él se beneficiaría también. (Análogamente, la creencia de que cierta conducta está prescrita por Dios y de que todas las desviaciones serán castigadas con calamidades podría ser una creencia útil para la gente, sea o no verdadera, o tenga o no sentido, siempre que garantice a otros la continuidad de la conducta de una persona.) Eso plantea la posibilidad de una explicación sociobiológica, no de patrones particulares de conducta, sino

de la creencia en un orden moral objetivo. La creencia en la *corrección* podría haber sido seleccionada. (¿Podría una creencia en principios *deontológicos* servir a una función interpersonal similar y haber sido, así, seleccionada?)

Si la gente ha de recibir garantías acerca de mi conducta futura, acaso no baste con que yo me limite a anunciar mis principios; quizá los otros necesiten ver, llegada la ocasión, que yo me adhiero realmente a esos principios. Sin embargo, quizá los principios que yo creo más correctos o adecuados resulten difíciles de observar en acto para otros; esos principios máximamente adecuados podrían responder a detalles contextuales sutiles, a matices de historia, o de motivación, o de relación, ignorados por los demás o difícilmente controlables por ellos. La justicia, se dice, no sólo debe hacerse, sino parecer que se hace. Sin embargo, ¿qué ocurriría si lo que puede ser visto y reconocido por otros es menos complejo que lo que la justicia (plenamente) adecuada requiere? La función interpersonal de asegurar a otros que se hace justicia o que se siguen los principios podría exigir el acatamiento de principios que son menos sutiles y matizados, pero cuyas aplicaciones (y malaplicaciones) pueden a veces ser controladas por otros.*

Así, puede haber un conflicto entre ajustar un principio a una situación y generar confianza pública a través de ese principio. Cuanto más ajustado el principio, menos fácilmente pueden otros controlar sus aplicaciones. Por otro lado, a partir de un umbral de ajuste, un principio puede dejar de inspirar confianza, no porque no sea susceptible de control, sino porque sus aplicaciones dejan de ser deseables. Se ha dicho —el asunto se presta a cierta controversia— que los juicios morales de las mujeres están más precisamente ajustados a los detalles y matices situacionales de relación y motivación que los de los hombres[14]. Esa diferencia, si realmente existe, podría ser explicada por el hecho estadístico de que las mujeres toman (o anticipan la toma) de decisiones con menos frecuencia en un ámbi-

* David Kreps, *A Course in Microeconomic Theory* (Princeton: Princeton University Press, 1990) [trad. cast.: *Curso de teoría microeconómica*, Madrid, McGraw-Hill, Interamericana de España, 1994], pág. 763, refiere que Robert Wilson sostiene que las empresas públicas de contabilidad que realizan auditorías externas de los negocios, con objeto de garantizar a los inversores potenciales que los auditores mismos no están sobornados por las empresas que auditan, deben seguir reglas de auditoría establecidas, reglas cuya aplicación pueda ser controlada externamente, aun si esas prácticas no proporcionan la información más reveladora sobre las finanzas de un negocio. Puesto que *puede* controlarse la aplicación de esas reglas establecidas, la empresa auditora es capaz de mantener su reputación como tercero independiente.

to no familiar en el que la base o los motivos de decisión son objeto de sospecha. Si hay que dar garantía a otros en un ámbito (público), cualquiera en ese ámbito puede necesitar atarse (de algún modo) a los dictados de lo que *puede* proporcionar garantías, y los principios constituyen un mecanismo de este tipo. Se han hecho predicciones acerca de los cambios morales que traería consigo el ingreso de una muchedumbre de mujeres en ámbitos previamente masculinos —un buen asunto por muchos motivos—, pero no está claro que lo que experimentara un gran cambio fueran los ámbitos y no las mujeres.

Los principios de otra persona me permiten predecir con razonable (aunque acaso no perfecta) exactitud algunos aspectos de su conducta, llevándome así a contar con esos aspectos. A esa otra persona, sin embargo, sus principios no le parecen primordialmente mecanismos de predicción. Sólo raramente trata la gente de *predecir* su propia conducta futura; normalmente, se limitan a *decidir* qué hacer. En cambio, los principios de una persona desempeñan un papel en la producción de esa conducta; la persona *orienta* su conducta sirviéndose de los principios. Mi conocimiento de sus principios afecta a mi estimación de la probabilidad de que se conduzca de una determinada manera, a mi estimación de la probabilidad de que se conduzca de esa manera. *Para ella,* los principios no afectan (meramente) a las estimaciones de las probabilidades, sino a esas mismas probabilidades; los principios no ponen de manifiesto el modo en que ella se comportará, sino que son mecanismos que contribuyen a determinar lo que va a (decidir) hacer.[15]

Las funciones personales

Debido a que los principios de conducta tienen una función personal (o una función intelectual), independiente de las cuestiones relacionadas con la interacción social, son capaces de desempeñar y cumplir su función interpersonal. (Podría bastarles a los demás con pensar —erradamente— que los principios cumplen alguna función personal para alguien.) Esa función interpersonal —garantizar a otros nuestra conducta frente a tentaciones, y así, inducirles a actuar coordinadamente con nuestras acciones— no podría aparecer (como solución de un juego de coordinación) ni mantenerse sin fundarse en la matriz personal. ¿Cuáles son, pues, las funciones personales e intrapersonales de los principios, y de qué modos las cumplen éstos?

Los principios pueden ser una de las vías por las que una persona puede definir su propia *identidad:* «Soy una persona con *estos* principios». Además, los principios seguidos en el transcurso de un período de tiempo son una de las vías por las que una persona puede integrar su vida a lo largo del tiempo y dotarla de mayor coherencia. Algunos podrían decir que es bueno tener principios porque es un modo de ser coherente. No obstante, si las acciones son (lógicamente) incoherentes en sí mismas o entre ellas —ir al cine tal día, y dejar de ir el mismo día—, entonces no será posible ejecutarlas todas, y no serán necesarios los principios para evitar la incoherencia. Entre las acciones cuya ejecución conjunta es lógicamente posible, la adhesión a principios no añade coherencia lógica adicional. Una acción *puede* ser incoherente con un principio, y por lo mismo —derivativamente—, con otras acciones que casan con ese principio. Pero si uno se limitara simplemente a querer evitar tal incoherencia, podría hacerlo sin tener *ninguna* clase de principios. Con todo, los principios prestan a las acciones cierta urdimbre. A su través, las acciones y la vida de uno pueden adquirir mayor coherencia, mayor unidad orgánica. Eso puede ser valioso en sí mismo.

¿Qué significa definirse a uno mismo o definir la propia identidad en términos de principios? ¿Deberíamos construir el yo como un sistema de principios? Éstos pueden incluir principios para transformar principios ya existentes y para integrar principios nuevos, principios, así pues, para alterar el yo también en términos de principios. (Si una persona viola sus principios, ¿amenazaría eso con destruir su yo?) Pero los fines continuados podrían integrar también la vida y las acciones de una persona a lo largo del tiempo. ¿Por qué definirse a sí mismo con principios y no con fines? Una persona que no se *define* a sí misma con principios podría, sin embargo, *tener* principios, no como un componente interno de su identidad, sino como una restricción externa puesta a las acciones de una identidad separada, distinguible. Piénsese en los temas kantianos de la autocreación y la autolegislación. Mas si los fines elegidos pueden llevar a la autocreación, ¿por qué es necesaria la auto*legislación*? ¿Depende ese papel desempeñado por los principios de las controvertidas tesis kantianas acerca de lo (único) que da lugar a la libertad autónoma?

Esas funciones personales de los principios tienen que ver con el conjunto de la vida o de la identidad de uno, o al menos con partes importantes de ese conjunto. Los principios funcionan también para una persona, más modestamente, al micronivel. Una función

intrapersonal de los principios morales está relacionada con nuestro compromiso con ellos. Cuando arrancamos proyectos a largo plazo, se plantea la cuestión de si nos atendremos a ellos en el futuro, de si —como dirían algunos— nuestros futuros yoes los ejecutarán. Sólo si la respuesta es sí, podría valer la pena empezar un proyecto particular, y empezarlo sólo sería racional si tenemos alguna garantía de que continuará. Si el que yo sostenga algo *como principio* ahora crea un coste mayor a una futura desviación del mismo —esa misma acción sería menos costosa si no hubiera desviaciones del principio—, entonces un proyecto que incorpora un principio para el presente y para el largo plazo será un proyecto que estaré menos inclinado a abandonar, no porque tenga yo algún principio adicional que me aferre a mi proyecto, sino porque ese proyecto incorpora un principio que yo (probablemente) seguiré albergando. Así como los principios cumplen una función interpersonal de dar garantías al otro —que puede contar con mi conducta a la hora de planear la suya—, también tienen la función intrapersonal de permitirme a mí contar con cierta conducta de mi yo futuro —dado que éste probablemente albergará el mismo principio—. Por ende, puedo razonablemente emprender algunos proyectos cuya deseabilidad depende de cierta conducta futura mía.

Dentro del proceso de toma de decisiones de una persona, los principios podrían funcionar como un mecanismo exclusionario o de filtro: en situaciones de elección, descarte usted las opciones que impliquen acciones que violan sus principios. Así, los principios le ahorrarían a una criatura de «racionalidad limitada» esfuerzo decisor y tiempo de cálculo. Sin embargo, no es necesario que la exclusión sea absoluta: si no hallamos ninguna acción suficientemente buena (por encima de cierto nivel de aspiración) entre las opciones disponibles, podríamos reconsiderar una acción previamente excluida.

VENCER LA TENTACIÓN

La función intrapersonal central de los principios en la que quisiera centrarme ahora es la de dejar atrás tentaciones, obstáculos, distracciones y diversiones. El psicólogo George Ainslie ha presentado una teoría para explicar por qué nos embarcamos en una conducta impulsiva, que sabemos contraria a nuestros intereses a largo plazo, y los mecanismos que usamos para lidiar con las tentaciones a que está expuesta nuestra conducta.[16] Antes de entrar en el trabajo de Ainslie, será útil poner algunos cimientos.

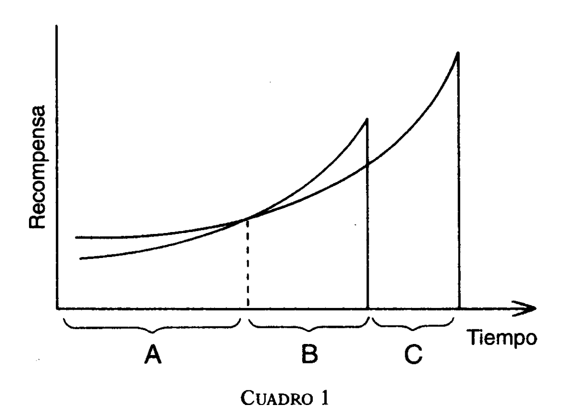
Los datos económicos y psicológicos muestran que ahora nos preocupamos menos por una recompensa futura de lo que lo haremos luego cuando la recompensa vaya a ser efectiva: «descontamos» el futuro. La utilidad que para nosotros tiene en el presente recibir una recompensa futura es menor que la utilidad que nos deparará la recompensa cuando la recibamos, y cuanto más distante la recompensa, menor la utilidad en el presente. Se trata de un fenómeno interesante por sí mismo, y podemos cuestionar su racionalidad. En nuestros planes y proyectos de acción, ¿deberíamos valorar siempre, en todos los momentos, una recompensa igual como la valoramos cuando la recibimos? Es verdad que también queremos tomar en cuenta la incertidumbre de que sobrevivamos hasta el momento de recibir la recompensa, o la incertidumbre de que de verdad la recibamos —es posible que no haya certeza sobre ningún acontecimiento—. En nuestros cálculos presentes, pues, queremos usar un valor esperado, descontando el valor de la recompensa futura por su probabilidad. Mas, ¿no debería permanecer constante la utilidad de recibir realmente la recompensa, independientemente del momento temporal?

La preferencia temporal —el término empleado por algunos economistas para referirse a la utilidad descontando el futuro— quizá sea la vía transitada por la evolución para inculcar en criaturas que no pueden ejecutar tales cálculos probabilísticos anticipatorios un mecanismo que tenga un efecto aproximadamente igual. La preferencia temporal innata quizá constituya una suerte de regla tentativa que nos acerca a la conducta o a las decisiones que habrían resultado de cálculos previos, al menos por lo que hace a las recompensas (y castigos) relacionadas con la adaptación inclusiva; es posible que esa preferencia temporal haya sido seleccionada.[17] A los seres dotados de un aparato cognitivo capaz de tomar explícitamente en cuenta las incertidumbres relacionadas con una recompensa futura y de ejecutar explícitamente una operación de descuento del futuro se les plantea, entonces, un problema. Si ya está ínsita en nosotros una preferencia temporal innata —el intento evolucionario de ejecutar para nuestros ancestros el descuento probabilístico— y si, además, lo que descontamos explícitamente en nuestros cálculos probabilísticos es el valor presente de la recompensa futura (ya descontado a través de la preferencia temporal), entonces lo que ocurre es un *descuento doble*. Y sin duda eso es demasiado. Parecería que los seres que son suficientemente sofisticados como para percibir todo esto y que ejecutan cálculos de valor esperado deberían usar estimaciones actuales de la utilidad que tendrán las recompensas futu-

ras cuando se reciban (las cuales son, entonces, explícitamente descontadas por las probabilidades), en vez de los valores actuales, descontados e influidos por la preferencia temporal, de esas recompensas futuras. De otro modo, deberían saltarse los cálculos del valor esperado y mantenerse en las preferencias temporales evolucionariamente inculcadas.[18] Sin embargo, si la preferencia temporal pura es un fenómeno racional en sí mismo, no *simplemente* un substituto evolucionario del descuento probabilístico, y si tal modelación evolucionaria hubiera tenido lugar, entonces la situación es más complicada.

Las curvas que describen el descuento típico de la preferencia temporal no necesariamente son líneas rectas o exponenciales; pueden ser curvas hiperbólicas.[19] Ainslie observó que dos curvas de este tipo marcadamente cóncavas (como las hiperbólicas) pueden cruzarse, y sacó las conclusiones de ese hecho. (En el cuadro 1, la utilidad de una recompensa se mide en el eje de las y; su utilidad para una persona en un momento dado se mide por la altura de su curva en ese momento. La curva tiene una pendiente descendiente hacia la izquierda porque una recompensa futura tiene menos valor antes.) Supongamos que hay dos proyectos o planes de acción que llevan a recompensas distintas, donde recibir la recompensa más temprana posible, la más pequeña de las dos, impedirá o estorbará la recepción de la recompensa mayor posterior. Una persona actúa a lo largo del tiempo aferrándose al proyecto que significa la mayor utilidad en ese momento. En el intervalo de tiempo A, la recompensa más lejana posee la mayor utilidad; en el intervalo de tiempo B, sin embargo, la recompensa más próxima posee la mayor utilidad. Puesto que la utilidad mayor puede cosecharse sólo al final del intervalo de tiempo C, la persona debe atravesar ese período intermedio B sin atender a la recompensa menor. Esto presenta un problema: durante ese período intermedio, la perspectiva de recibir esa recompensa menor *ya* tiene una utilidad mayor que la perspectiva de recibir después la recompensa mayor.

¿Por qué suponemos que la persona *debería* tratar de dejar atrás ese período de tiempo intermedio? ¿Por qué no debería coger la recompensa más pequeña, pero más inmediata?[20] ¿Qué es lo que hace que los períodos A y C, en los que la recompensa se vislumbra máxima, sean los períodos apropiados para decidir qué elección es la mejor? En esos períodos, la persona preferirá actuar para conseguir la mayor recompensa; durante el período B, preferirá actuar para conseguir la recompensa menor (es decir, una recompensa menor, cuando la obtiene, que la otra recompensa en el momento de obte-



CUADRO 1

ner*la*). ¿Dónde estamos *nosotros* cuando decimos que evitar la tentación es la mejor alternativa, y por qué es este punto de vista más apropiado que el de la persona dentro del intervalo temporal B?

He aquí una sugerencia. El intervalo temporal B no es un alto apropiado para decidir lo que debería hacer la persona porque B no es una muestra representativa de su punto de vista al respecto. Los intervalos temporales A y C componen, sumados, un intervalo temporal más dilatado. Además, cuando añadimos sus juicios *después* del momento de recepción de las recompensas y representamos en una gráfica qué recompensas le parecen *entonces* más grandes, hallamos que inmediatamente después de consumir la recompensa menor desea no haberlo hecho, mientras que tras consumir la recompensa mayor (al final del intervalo temporal C), continúa prefiriendo la elección de la recompensa mayor. Sugiero que, a menudo, lo que hace de resistir la tentación y de aceptar la recompensa mayor la opción preferida es que es la preferencia que alberga la persona la mayor parte del tiempo: es su preferencia (razonablemente estable; la otra es su preferencia en un momento no representativo.[21] (Dejando de lado cualesquiera preferencias-después-de-los-hechos, si el intervalo temporal B durara más que los intervalos A y C, ¿tendríamos tan claro *en ese caso* que hay que resistir la tentación?) Las tentaciones no siempre deberían ser resistidas; no deberían serlo cuando el deseo de una recompensa mayor (que incluye la preferencia-después-de-los-hechos) coincide con la preferencia personal del período más dilatado de tiempo. Apunto a ese criterio como un criterio rebatible, no concluyente. Tiene la virtud de andar cerca de las preferencias de una persona (aunque no está atado a una preferencia particular local), en vez de limitarse a decir que simplemente

está en el interés de la persona perseguir la recompensa mayor posterior (por causa de lo que esa misma recompensa representa) y, por lo tanto, resistir la tentación, o a decir que el criterio relevante es —y resistir la tentación sirve a— la maximización de la utilidad a lo largo de un ciclo vital completo.[22]

Ainslie describe varios mecanismos para hacer que uno mismo pase de largo por este período intermedio de tentación. Esos mecanismos incluyen: emprender una acción durante el intervalo temporal A que haga imposible la persecución de la recompensa menor durante el intervalo B (por ejemplo, la acción de Ulises de atarse al mástil); emprender una acción durante el intervalo A (hacer una apuesta con otra persona, quizá) que introduzca un castigo si ustedes aceptan la recompensa menor, modificando así la utilidad de *esa recompensa* durante el intervalo B; dar pasos en el intervalo A que impidan advertir o entretenerse en las virtudes de la recompensa menor durante el período B;[23] y, por último —lo que nos ocupa ahora—, formular un principio personal general de conducta.

Un principio general de conducta agrupa las acciones; clasifica con ellas una acción particular. Por ejemplo: «Nunca picotees entre comidas»; «No fumes nunca más otro cigarrillo». (Podría pensarse que los principios son más profundos y menos mecánicos que las reglas —ésa es al menos la distinción habitual en la filosofía del derecho—, pero a los efectos presentes ignoraré las distinciones de este tipo.) Podríamos tratar de representar el efecto de esta agrupación de acciones mediante principios en el marco de la teoría de la utilidad y de la decisión del modo que sigue. Al clasificar conjuntamente las acciones como acciones pertenecientes al tipo *T*, y al tratarlas simultáneamente, un principio vincula las utilidades de las acciones-*T* (o las utilidades de sus resultados). Sería demasiado fuerte decir que, debido al principio, todas las acciones-*T* deben tener la misma utilidad. Una acción-*T* particular puede caer también bajo tipos y principios distintos, en los que no cabrían otras acciones-*T*, motivo por el cual sus utilidades serían distintas. Lo que hace un principio es fijar una *correlación* entre las utilidades de las varias acciones que caen bajo él. Formulándolo en términos de preferencias: cuando acciones del tipo *T* se clasifican jerárquicamente con otras acciones en un orden de preferencias, habrá una correlación entre la jerarquía de órdenes de las acciones-*T*. Si, no obstante, esa correlación fuera el único efecto que tuviera en las utilidades de las acciones que caen bajo principios la adopción o la aceptación de los mismos, entonces los principios no servirían para dejar atrás las tentaciones.

La marca distintiva de un principio («Nunca picotees entre comidas»; «No fumes nunca más otro cigarrillo») es que vincula la decisión de acometer una acción particular inmediata (picotear *esta* tapa, fumar *este* cigarrillo) al entero conjunto de acciones englobadas por el principio. Esa acción vale ahora para todo el conjunto. Adoptar el conjunto es como si ustedes hicieran verdadero lo siguiente: si ustedes emprenden esta particular acción perteneciente al conjunto, emprenden todas las acciones del conjunto. Lo que ahora está en juego es mucho mayor. Vincular la utilidad de esta acción de picotear a la desutilidad de todas las acciones futuras de picotear puede ayudarles a ustedes a dejar atrás las tentaciones del período B; se les ha modificado a ustedes ahora la utilidad de este picoteo particular. Este picoteo vale para todos los picoteos, y en este momento previo, la utilidad presente de estar delgado o sano más adelante rebasa con mucho la utilidad presente de esos placeres gastronómicos distantes; la desutilidad presente de una mala salud o de una mala figura se convierte en un rasgo de la particular acción de picotear tal como se contempla en el presente.[24]

Mas ¿por qué presumir que la persona formulará un principio durante el período temporal A y no en el período B? ¿Por qué no habría la persona de picotear ahora y formular un principio que le recomendara picotear todo el tiempo, o, más generalizadoramente, un principio que le recomendara ceder siempre a la tentación inmediata? Pero formular y aceptar un principio así (junto a la acción de picotear ahora) no traería por sí mismo ni la recompensa inmediata ni la maximización de las recompensas a lo largo del tiempo. Lo que hace, generalmente, es reducir la demora de la recompensa. Durante el período B, empero, expuesto a una tentación particular, ¿deseo *siempre* reducir la demora de todas y cada una de las recompensas? No. Pues, aun cuando, con respecto a una recompensa particular, estoy en el período B, con respecto a otras muchas (parejas de) recompensas, estoy en el período A (o en el período C). Con respecto a estos otros y más distantes pares de recompensas menores y mayores, yo no deseo ahora recibir la recompensa más inmediata, aunque sí deseo ahora recibir una *particular* recompensa que es más inmediata porque estoy en *su* período B. Puesto que las tentaciones están dispersadas a lo largo del tiempo, en cualquier *momento* nos hallamos más en los períodos A o C que en el B. De aquí que no podamos aceptar un principio que nos recomendara sucumbir siempre a la tentación.*

* Quien propusiera sucumbir a la tentación podría replicar: «Usted dice que no siempre *queremos* sucumbir a la tentación. Pero usted dice que un principio es un

Al adoptar un principio, hacemos valer una acción por muchas otras, alterando así la utilidad o desutilidad de esa particular acción. Tal modificación de utilidades es resultado del ejercicio de nuestras facultades y capacidades para hacer que una acción *valga por* o *simbolice* a otras. Violar el principio esta vez no implica necesariamente violarlo siempre: picotear ahora no implica que nos convirtamos necesariamente en perpetuos picoteadores. Antes de adoptar el principio no era verdad que realizar esta acción ahora implicaría hacerla siempre. La adopción del principio forja esa conexión, de manera que el castigo por violar el principio esta vez se convierte en la desutilidad de violarlo siempre. Sería instructivo investigar *cómo* somos precisamente capaces de hacer eso.

El hecho de que *podamos* tiene importantes consecuencias. Podemos modificar así nuestras utilidades (adoptando un principio y haciendo que una acción valga por otras), pero no podemos hacerlo con demasiada frecuencia y atenernos a ello. Si violamos un particular principio que hemos adoptado, no tenemos razones para esperar que la siguiente ocasión será de algún modo distinta. Si cada ocasión es la misma, y lo hacemos esta vez, ¿no lo haremos siempre que se presenten tales ocasiones? A menos que podamos distinguir esta ocasión de las siguientes, cargándonos así de razones para creer que *más adelante* conferiremos peso a esa distinción, de modo que no nos dispensaremos una vez más formulando otra distinción a la que de nuevo no querremos adherirnos más adelante, ejecutar la acción esta vez nos llevará a esperar que continuaremos ejecutándola en el futuro. (Formular una distinción que permite esta acción ahora excluyendo futuras repeticiones es, sin embargo, formular otro principio; debemos tener razones para pensar que nos adheriremos más a uno que a otro, o la reformulación no hará creíble nuestra abstención futura.) Realizar la acción esta vez, en *esta* situación, significa que continuaremos haciéndolo en el futuro. ¿No basta eso para modificar la utilidad presente de hacerlo esta vez, cargando a esta

---

mecanismo para dejar atrás nuestros posibles deseos presentes. De modo que quizá necesitemos un principio para dejar atrás el deseo de no sucumbir siempre a la tentación». Aparte del hecho de que esta réplica bordearía la paradoja, un principio se adopta (del modo más fácil) durante el período *t*, cuando un deseo contrario es más fuerte de lo que lo es la tentación mientras dura *t*. (La tentación cobrará plena fuerza sólo después de *t*.) Y no habrá un período de tiempo en el que el deseo de sucumbir *siempre* no sea más débil que un deseo contrario. (O, si surgiera un período temporal tal, cualquier principio adoptado pronto sería derrotado por otro deseo que no sería meramente momentáneo.)

particular acción presente la desutilidad de todas sus repeticiones futuras?

La expectativa es que si lo hacemos ahora, lo haremos repetidamente en el futuro. Pero, ¿el hacerlo ahora una vez *afectará* al futuro, *hará* más probable que repitamos la acción? ¿U ocurre simplemente que esta acción afecta a nuestra *estimación* de la probabilidad de que haya repeticiones? Hay dos situaciones que considerar. Si no se ha adoptado previamente ningún principio que excluyera la acción, realizar la acción ahora quizás tenga un efecto menor, de acuerdo con la ley psicológica «del efecto», en la probabilidad de repetición: el refuerzo positivo de una acción aumenta la probabilidad de que se repita en el futuro. Y la estimación de la probabilidad de repetición quizás aumente un poco si esa acción viene a añadirse a unas cuantas acciones similares en el pasado. En cambio, si se había adoptado previamente un principio, actuar violando el principio aumentará la estimación del observador, y la propia del agente, de la probabilidad de que éste repita su particular acción. El principio se ha derrumbado; ha desaparecido un obstáculo a la acción. Por lo demás, advertir eso puede resultar desalentador para el agente e inhibir su esfuerzo para evitar la acción en el futuro. (Obsérvese que una acción que afecta a la *estimación, por parte del agente,* de la probabilidad de acciones similares futuras puede generar desaliento y, por consecuencia, afectar a la probabilidad real de repetición.) Formular un principio que constituya un obstáculo adicional a las acciones que excluye es un modo de vincular realmente los efectos de todas ellas a los efectos de cualquiera (previa) de ellas. Cuanto más ha invertido uno en un principio, cuanto mayor empeño previo ha puesto en adherirse a él, mayor es el coste de violarlo ahora. (Pues ¿cuál es la probabilidad de que ustedes continúen adhiriéndose en el futuro a otro principio si no consiguieron atenerse a éste después de tanto esfuerzo?) Además, adherirse al principio esta vez es un tipo de acción sometido a la ley del efecto: el refuerzo positivo hace más probable que esta adhesión a este principio se dé en el futuro.

Los efectos de violar un principio pueden ser aún más generales, pues la probabilidad o la credibilidad que tenga el que ustedes se adhieran con éxito a *algún* principio en *algún* ámbito (expuestos a una tentación tan fuerte como la que les hizo sucumbir esta vez) puede verse afectada. Es verdad que ustedes pueden tratar de ceñir y limitar el peligro *a ese ámbito,* pero aquí se les presentará un problema idéntico —un nivel más arriba— al de tratar de limitar el daño *dentro* de ese ámbito a sólo *esta* acción violadora. Los principios

deontológicos quizá consigan su mayor fuerza cuando su violación amenaza directamente *cualquier* acción futura guiada por principios: si yo violo *este* principio (en esta circunstancia), ¿cómo puedo creer que alguna vez tendré éxito en la adhesión a cualquier principio (deseable)? En un exceso de celo kantiano, alguien podría tratar de incrementar el efecto potencial de dispersar el desastre formulando un (meta-)principio, según el cual nunca podría violarse ningún principio. Mas, aun cuando hacer valer una violación por todas podría rebajar la probabilidad de cualquiera de ellas, las consecuencias de la menor violación quedarían peligrosamente magnificadas. Eso no quiere decir que la violación de un principio, dado que una acción vale por todas, quite a tal punto mordiente a un principio que una persona pueda violarlo libre e impunemente. Una acción tiene la desutilidad de todas, pero también la siguiente, aun hecha ya la primera acción. Esa desutilidad puede ser eludida cancelando el principio, no violándolo; pero entonces se enfrenta uno a la misma desutilidad que trataba de evitar adoptando un principio.

Puesto que la adopción de un principio es ella misma una acción que afecta a los vínculos de probabilidad que se dan entre otras acciones, resulta adecuado poner algún cuidado en la elección de los principios que se adoptan. Hay que considerar no sólo los posibles beneficios resultantes de la adhesión, sino también la probabilidad de violación y los efectos futuros que ésta traería consigo. Podría ser mejor adoptar un principio menos bueno (cuando se secundara) pero más fácil de mantener, sobre todo si ese principio no siempre pudiera estar disponible como un rellano creíble en el que sostenerse si fallara la adhesión al principio más estricto. (También: se desean principios lo suficientemente perfilados como para que sus violaciones destaquen con claridad, de manera que los yoes futuros no puedan sortear fácilmente la cuestión de si el principio es secundado.) Sin duda podría formularse una teoría de la elección óptima de principios teniendo en cuenta consideraciones de este tipo.[25]

Lo típico de un principio es referirse a todas las acciones de un grupo, y hacer valer por todas cada acción presente. Para cumplir su función de dejar atrás tentaciones como las del período B, tiene que referirse a *todas* las acciones de un cierto tipo. No tenemos principios que digan que *la mayoría* de **P**s deberían ser **Q**s, o que el 15 % de **P**s han de ser **Q**s. (O, si los tenemos, entonces no están pensados para lidiar con las mismas tentaciones.) Algunas veces, sin embargo, todo lo que necesitamos es realizar alguna acción durante algún tiempo, o la mayor parte del tiempo (por ejemplo, saltarse los postres la mayoría de las noches, pagar la mayoría de las facturas cada

mes). El modo en que nos servimos de los principios para conseguirlo consiste, no obstante, en formular un enunciado que hable de «todos» o «cada» y que, a pesar de ello, sea coextensivo con la mezcla que deseamos. *Cada* mes, pague usted la mayoría de sus facturas; *todas* las semanas, sáltese los postres la mayoría de las noches; *cada* año vaya a algún claustro general de facultad. Un profesor —no yo mismo— cuyo principio sea no dar muchos excelentes de promedio en *cada* clase. Así, cada semana, o mes, o clase, vale por todos. Eso explica por qué los principios que sirven para vencer las tentaciones afectan a todos los miembros de un conjunto, no sólo a algunos. (Una norma podría limitarse a un n por ciento, no siendo n ni 0 ni 100, pero un principio, no.) Un principio tiene ciertas funciones, y para cumplirlas, un caso debe valer por o simbolizarlos todos. El carácter totalizante de los principios viene, pues, en apoyo de nuestra idea de las funciones que tienen los principios y de las vías por las que las cumplen.[26]

Acaso parezcan los principios rudos mecanismos para cumplimiento de nuestros fines; su espectro universal —renunciar a *todos* los postres, a *todas* las diversiones hasta que se haya realizado la tarea— quizá contenga más de lo necesario para conseguir el fin. El margen de libertad en lo que haya que considerar cubierto por el «todos» (postres, semanas) mitiga esto un poco, restringiendo los excesos de los principios. Con todo, subsistirán algunos inconvenientes. Si hubiera un umbral claro de n repeticiones de una acción, pasado el cual las consecuencias de continuar esa acción perturbaran el fin, pero antes de llegar al cual el fin pudiera aún ser alcanzado, ¿no ejecutarían las personas racionales n veces la acción para abstenerse luego de ella? (Se necesitaría una formulación más complicada si cada repetición incrementara la dificultad de conseguir el fin.) No sería necesario ningún principio para excluir la acción $n + 1$, puesto que la acción misma daría un mal saldo de consecuencias. Eso podría ser una teoría acerca del momento (aproximado) en que una persona decide dejar de fumar (o de ganar peso, etc.) y, por lo tanto, acerca del momento en que decide instituir un principio. Sin embargo, dada la tentación, es un principio que necesita ser instituido *entonces*.

COSTES SUMERGIDOS

Un método mencionado por Ainslie para dejar atrás el intervalo tentador B es éste: *comprométase* en el intervalo previo A a buscar

la recompensa mayor en C y en B. Una modalidad de ese compromiso es invertir, mientras dura A, muchos recursos en la persecución (futura) de esa mayor recompensa. Si yo creo que sería una cosa buena para mí ver mucho teatro o asistir a muchos conciertos este año, y si sé que llegada la noche del espectáculo con frecuencia no me siento motivado para salir, entonces puedo comprar por anticipado entradas para asistir a muchas sesiones, aun a sabiendas de que habrá entradas a la venta la misma velada del espectáculo. Puesto que no querré despilfarrar el dinero ya gastado en las entradas, asistiré a más espectáculos que si dejara la decisión de asistir a la noche misma en que se ofrecen. Es verdad que quizá no use *todas* las entradas —el letargo indolente puede vencerme algunas noches—, pero asistiré con más frecuencia que si no hubiera hecho la compra anticipada de entradas. Sabiendo todo esto, compraré anticipadamente las entradas para moverme a mí mismo a la asistencia.

Los economistas presentan una doctrina, según la cual toda toma de decisión debería limitarse a prestar atención a las consecuencias (presentes y) futuras de varias acciones alternativas. Ya se ha incurrido en los costes de las inversiones pasadas en esos cursos de acción. Aunque los recursos existentes pueden afectar a las consecuencias de las varias líneas de acción que ahora tengo abiertas —poseyendo ya la entrada, puedo asistir al espectáculo sin necesidad de un desembolso adicional futuro—, y por lo tanto tomarse en cuenta a través de esas consecuencias, el mero hecho de que los costes de promover cierto proyecto ya se han producido no debería pesar en absoluto a la hora de que una persona tome una decisión. Esos costes, los «costes sumergidos», como los llaman los economistas, son cosa del pasado; todo lo que cuenta ahora es el flujo de los beneficios futuros. Así, sentado esta noche en mi casa, si yo ahora prefiriera quedarme a salir y asistir a un espectáculo (sin tener que desembolsar nada), entonces quedarme en casa tendría mayor utilidad para mí que desplazarme y asistir al concierto; por lo tanto, debería quedarme en casa. En nada cambia las cosas el que yo haya gastado dinero en la entrada para el espectáculo —así razona la teoría económica, para la cual los «costes sumergidos» deben ser ignorados—.[27]

Puede que ésta sea una buena regla para la maximización de los beneficios monetarios, pero, por razones cotidianas, no constituye un principio general adecuado de decisión. *No* consideramos nuestros compromisos pasados con otros como irrelevantes salvo en el caso de que afecten a nuestros rendimientos futuros, como cuando romper un compromiso puede afectar a la confianza depositada en

nosotros por otros y, por lo tanto, a nuestra capacidad para lograr otros beneficios futuros; y *no* consideramos los esfuerzos que en el pasado hemos consagrado a proyectos, de trabajo o de vida, tanto en curso como irrelevantes (salvo en el caso de que su continuación haga más probable la consecución de beneficios de lo que lo harían otros proyectos recién iniciados). Pues esos proyectos nos ayudan a definir el sentido de nosotros mismos y de nuestras vidas.[28]

El particular asunto que hemos estado discutiendo sugiere, sin embargo, otro defecto en la doctrina que recomienda ignorar los costes sumergidos como un principio general de decisión. El hecho de que no ignoremos los costes sumergidos proporciona un modo de dejar atrás la tentación en el período B de elegir la recompensa menor pero más inmediata. Previamente, en el período A, cuando podemos ver claramente los beneficios de la recompensa mayor pero más distante, podemos sumergir los recursos y esforzarnos en conseguir esa recompensa a sabiendas de que, cuando llegue el tiempo de la tentación, el hecho de que no queremos (y de que no querremos) haber despilfarrado esos recursos contará para nosotros como una razón contraria a elegir la recompensa menor, sumándose a su desutilidad. Si yo sé que en una noche venidera me sentiré tentado por la recompensa inmediata menor de la comodidad (no tener que salir con lluvia, etc.), pero también sé que ahora y en el futuro estaré encantado de haber asistido a todos esos espectáculos, entonces puedo comprar las entradas ahora, por anticipado, para incitarme a mí mismo a renunciar a la acción de quedarme en casa cuando llegue la ocasión.

Todo el mundo ve como un problema, como una irracionalidad, o como una miopía indeseable, el sucumbir a la tentación de la recompensa menor durante el intervalo temporal B. La persona misma lo ve así —antes y después, si no en el momento apropiado— y no podemos dejar de verlo así cuando pensamos en ello. Los economistas consideran también irracional e indeseable otro tipo de conducta, el respeto de los costes sumergidos. Pero ahora vemos que esa conducta ulterior, previamente anticipada, puede servir para limitar y moderar el primer tipo de conducta indeseable (sucumbir a la recompensa menor pero más cercana). Podemos utilizar conscientemente nuestra tendencia a tomar en serio los costes sumergidos como un medio para incrementar nuestras recompensas *futuras*. Si esa tendencia es irracional, lo cierto es que puede utilizarse racionalmente para moderar y vencer otra irracionalidad. Si alguien nos ofreciera una píldora, tras ingerir la cual *nunca más* respetaríamos los costes sumergidos, andaríamos mal aconsejados si la

aceptáramos; nos privaría de una herramienta valiosa para dejar atrás tentaciones del momento (futuro). (¿Es posible que esta tendencia a respetar los costes sumergidos, que puede tener valor adaptativo, haya sido seleccionada por el proceso evolucionario?) Puesto que a veces tomar en cuenta los costes sumergidos es deseable (de modo que la condena general de los economistas está equivocada), y a veces, no, la deseabilidad de ingerir la píldora dependería de los números comparativos de, e iría de consumo con, estos dos tipos de situaciones encontradas.

Ya mencioné antes que cuanto mayor empeño se ha puesto en la adhesión a un principio diseñado para dejar atrás las tentaciones del momento, mayor es el coste de violarlo ahora. Es improbable que ustedes consigan atenerse a otro principio si no pudieron atenerse a éste a pesar del mucho esfuerzo previo dedicado a ello. Apercibirse de eso les da a ustedes buenas razones para atenerse a este principio —es el único salvavidas a la vista—, y por lo tanto, confiere mucho peso a la inviolabilidad del mismo en el caso de una tentación particular. Las agrupaciones de acciones (con el propósito de evitar tentaciones inmediatas) en cuyo secundamiento hemos tenido éxito ganan, por lo mismo, en tenacidad. Obsérvese que esto implica un fenómeno de costes sumergidos. Mi argumentación en favor de atenerme a *este* principio y al agrupamiento de acciones que va con él corre pareja de la siguiente reflexión: si no puedo atenerme a él a pesar de tanto esfuerzo previo, ¿cómo podría esperar atenerme a otro? Sólo seré capaz de argumentar así si soy alguien que respeta los costes sumergidos; sólo quien respeta los costes sumergidos podría tener una razón para adherirse ahora a este principio actual con objeto de eludir la tentación en vez de sucumbir a ella esta vez, para luego formular un principio diferente que a su tiempo acabará cayendo también, quizás a la primera prueba. Lo que da margen para el mantenimiento de *este* principio son los costes sumergidos. (No se diga que éstas son consideraciones orientadas al futuro acerca de las consecuencias venideras de dos distintos cursos de acción —atenerse a la política presente, en vez de sucumbir a la tentación, para luego formular una nueva política— y que, en consecuencia, la persona que no respeta los costes sumergidos puede seguir la misma línea de razonamiento; pues sólo merced a la conocida tendencia a respetar los costes sumergidos tendrá —y se verá que tiene— un curso de acción consecuencias significativamente distintas de las de otro. ¿Por qué habríamos de pensar, si no, que es menos probable que yo secunde el nuevo principio después de violar el viejo si yo no lo estoy violando ahora?) ¿Podría el

cotidiano fenómeno del respeto de los costes sumergidos jugar algún papel en nuestro secundamiento de principios que ya hemos adoptado? Sabemos ahora que si podemos arreglárnoslas para adherirnos a este principio por algún tiempo, el hecho de que hayamos invertido en él nos suministrará en el futuro, respetuosos con los costes sumergidos como somos, razones para continuar secundando el principio entonces —y esto puede proporcionarnos ahora alguna razón—.[29]

A la vista de esas funciones cumplidas por nuestro respeto de los costes sumergidos, el economista podría replicar que, para una persona perfectamente racional en todo lo demás, el respeto de los costes sumergidos no es en absoluto deseable; sólo lo necesitaría alguien con alguna *otra* irracionalidad. Tal conclusión no es tan evidente, sin embargo, aun dejando de lado lo que antes se mencionó; los compromisos con otras personas y la inversión pasada en nuestros proyectos laborales y vitales. Pues podría resultar interpersonalmente útil tener un medio para convencer a otros de que nos atendremos a proyectos o a objetivos aun en el caso de amenazas que aparentemente hacen militar esa adhesión en desfavor de nuestro futuro —como un modo de de potenciar el anuncio o la ejecución de esas amenazas—.[30] Esto podría resultar útil aun en el caso de que ustedes no tuvieran ninguna otra tendencia a la conducta irracional ni tampoco la tuvieran aquellos a quienes ustedes tratan de convencer.[31] No obstante, el tema de contrarrestar o mantener a raya una irracionalidad sirviéndose de otra es digno de ser destacado. ¿Se pueden poner conscientemente otras cosas que creemos irracionales —la debilidad de la voluntad, el autoengaño, o las falacias argumentativas, pongamos por caso— al servicio de la frustración o de la limitación de otras irracionalidades o acontecimientos indeseables? (¿Podría acaso el conjunto entero de esas tendencias aparentemente irracionales funcionar incluso mejor que el *entero* conjunto de tendencias aparentemente —consideradas por separado— racionales?)

Permítaseme mencionar otra técnica que podría usar una persona para ayudarse a sí misma a superar el período tentador B, en el que la recompensa menor parece tan grande. Esa persona podría considerar qué acción recomendaría ella misma a otra persona (cuyo bienestar le importara) en la misma situación para, luego, seguir ese consejo ella misma. Distanciarse uno mismo de la situación, mirar el panorama impersonalmente, no simplemente mirarlo desde una perspectiva temporal determinada, podría ser una vía para eludir la atracción de una recompensa ciertamente cercana (pero en defi-

nitiva, pequeña). Tal procedimiento requiere una capacidad para contemplar impersonalmente una situación en la que ustedes se hallan y para pensar que el mismo principio de elección que vale para otros debería valer para ustedes, que ustedes deberían emprender la misma acción que deberían emprender otros en esa situación. Tener una fuerte predisposición a una actitud imparcial de este tipo resultaría extremadamente útil para remontar el intervalo B en el que se cruzan las curvas y, por lo tanto, para maximizar la recompensa total de una persona. Y esa misma disposición constituye un componente del juicio ético; aplicar los mismos principios a la conducta propia y a la ajena.

No he mencionado hasta ahora una función de los principios: *trazar la línea.* Los principios marcan un límite que no se puede rebasar —«aquí trazo yo la línea»—. Pensamos: «si no la trazo aquí, ¿dónde la *habré* de trazar?». En un gradiente de situaciones, puede que no haya ningún otro lugar obvio, ningún lugar obvio en el territorio aceptable. (O quizás haya otro lugar aceptable, pero no nos sentimos capaces de trazar allí con éxito la línea.) Esto está conectado con la función de los principios, anteriormente mencionada, de dejar atrás la tentación del momento. En este caso, sin embargo, no se trata de una tentación, sino más bien de *un razonamiento* sobre el momento que es necesario dejar atrás. Si alcanzo *este* punto, argüiré que no hay razón especial para detenerse precisamente entonces, sería mejor haberme detenido mucho antes, allí donde *hay* una línea clara, una línea *especial.**

Creo que esto es lo que permite a los principios definir a una persona. «*Ésas* son las líneas que yo he trazado». Esas líneas lo perfilan e iluminan. Son sus límites externos. Ello es que una persona en circunstancias muy afortunadas, que sabe que no puede deslizarse realmente muy lejos por un gradiente indeseable, quizá pueda prescindir de trazar líneas específicas. En este sentido, pues, acaso

* La teoría de los juegos de coordinación de Thomas Schelling podría incorporar útilmente esta noción de «especial». Al tratar de coordinarme con otro, estoy buscando una acción de la que ambos podamos llegar a pensar que es especial (también deseable), de cuyo carácter especial ambos podamos llegar a apercibirnos —no simplemente llamativa, sino especial—. Cuando hay diez alternativas, nueve de ellas extremadamente llamativas, la alternativa especial podría ser precisamente la que no es nada llamativa —al menos a primera vista—.

He aquí un problema de coordinación. Cada uno de nosotros tiene que señalar un filósofo alemán de una determinada época, y si todos coincidimos en el mismo, cada uno de nosotros recibirá un gran premio. Los filósofos entre los que hay que escoger son Kant, Hegel, Fichte, Schelling y Jacobi. ¿Cuál escogerían ustedes?

no esté tan bien definido como alguien en circunstancias menos afortunadas.

LA UTILIDAD SIMBÓLICA

Hemos dicho que, tras adoptar un principio, realizar esta vez, en esta situación, ahora, una particular acción miope significa que continuaremos haciéndolo en el futuro. Esta acción *vale por* todas las demás acciones excluidas por el principio; realizarla es un *símbolo* de que también se realizan las demás. Ese hecho de *significar, valer por* y *simbolizar,* ¿está constituido por el entreveramiento de las dos hebras de conexión, ya discutidas, entre realizar la acción ahora y repetirla en el futuro? (Realizarla ahora afecta a su estimación de la probabilidad de que ustedes vuelvan a realizarla, y realizarla ahora modifica realmente la probabilidad misma de realizarla en el futuro.) ¿O es, acaso, la simbolización un hecho adicional, no agotado por esas dos hebras, sino un hecho que afecta a la utilidad de las acciones y los resultados alternativos? Yo creo que la simbolización es una importante hebra adicional, una hebra que debe tratar explícitamente una teoría adecuada de la decisión.

La teoría freudiana explica la ocurrencia y la persistencia de las acciones o los síntomas neuróticos a partir de su significado simbólico. Generadoras de consecuencias manifiestamente malas, esas acciones y esos síntomas aparentemente irracionales tienen un significado simbólico que no resulta obvio; simbolizan alguna otra cosa, llamémosle *M*. Con todo, la mera posesión de un significado simbólico no puede explicar por sí sola la ocurrencia o la persistencia de una acción o de un síntoma. Tenemos que añadir que aquello que esas acciones y esos síntomas simbolizan —es decir, *M*— tiene por sí mismo algún valor o utilidad (o, en el caso de que se trata de conductas que evitan, un valor negativo o una desutilidad) para la persona; y además, que esa utilidad de *M* que es objeto de simbolización se imputa retroactivamente a la acción o al síntoma, confiriéndo*les* así mayor utilidad que la que aparentan tener. Sólo así puede explicar el significado simbólico de una acción o de un síntoma por qué se eligió o por qué se manifestó. La teoría freudiana está obligada a sostener no sólo que las acciones y los resultados de éstas pueden simbolizar para una persona otros acontecimientos distintos, sino que pueden cargar ellas mismas con el significado emocional (y los valores de utilidad) de esos otros acontecimientos. Poseyendo un valor simbólico, las acciones se consideran como si tuvieran la utili-

dad de lo que ellas simbólicamente significan; la adhesión a un síntoma neurótico se produce con una tenacidad apropiada a aquello para lo que ella vale. (No tengo conocimiento de una formulación clara, por parte de la literatura freudiana, de esta ecuación, ni de la tesis, más débil, de que *parte* de la utilidad del objeto de simbolización se imputa retroactivamente al símbolo. De todas formas, creo que alguna versión de este tipo se encuentra implícitamente presupuesta en algunas explicaciones freudianas.) Las respuestas emocionales desproporcionadas a un acontecimiento real pueden sugerir que éste vale por otros acontecimientos u ocasiones para los que las emociones resultarían más adecuadas.[32]

Para que la acción simbólica se realice, tiene que tener de algún modo una utilidad mayor, un número mayor que representa el maximando, que las otras acciones a disposición del agente.[33] Ya he sugerido cómo ocurre esto. La acción (o uno de sus resultados) simboliza cierta situación, y la utilidad de esta situación simbolizada se imputa retroactivamente, a través del vínculo simbólico, a la acción misma. Obsérvese que la teoría estandard de la decisión cree también en una imputación retroactiva de utilidad, una imputación vehiculada por un vínculo causal (probabilístico). En virtud de garantizar el advenimiento de una situación particular, una acción llega a tener —se le imputan— las utilidades de esa situación en forma de utilidad esperada. Lo que añade el punto de vista que estamos discutiendo aquí es que la utilidad puede fluir hacia atrás, ser retroactivamente imputada, no sólo por la vía de las conexiones causales, sino también por la de las simbólicas.

Un indicador de que se da una conexión simbólica entre la acción y un resultado que juega un papel central en la decisión de realizarla, no sólo la conexión aparentemente causal (estoy pensando en casos en los que el agente no cree que la acción sea intrínsecamente deseable o valiosa), es la persistencia de la acción aun habiendo evidencia de que esa acción no tiene las consecuencias causales que se le presumen. A veces, una persona se negará incluso a atender o a tolerar esa evidencia, o alguna otra, de las consecuencias perniciosas de la acción o de la política emprendida. (Fundándose en ello, se puede decir que ciertas medidas tendentes a combatir la droga *simbolizan* la reducción de la cantidad de droga consumida y que las leyes de salario mínimo *simbolizan* la ayuda a los pobres.) A un reformador que quisiera evitar esas consecuencias perniciosas podría parecerle necesario proponer otra política (sin tales consecuencias) que simbolizara también efectivamente la actuación tendente a conseguir, o la consecución misma, del fin. Limitarse a parar

la acción en curso privaría a la gente de la utilidad simbólica de la misma, algo que no están dispuestos a aceptar.

Evidentemente, conferir un significado simbólico particular a una acción A trae consigo consecuencias causales en la medida en que afecta al tipo de acciones que ejecutamos, y una teoría puramente consecuencialista puede decir algo al respecto. Puede referirse a si conferir tal significado simbólico (o, más adelante, abstenerse de extinguirlo) constituye por sí misma una acción causalmente óptima. Pero este punto de vista será diferente del de una teoría puramente consecuencialista (no simbólica) de la acción A como tal, y no implica que debamos valorar la acción de eliminar o tolerar el significado simbólico únicamente por sus consecuencias causales.

Puesto que las acciones simbólicas son a menudo acciones *expresivas,* podrían también entenderse de modo diferente así: la conexión simbólica entre una acción y una situación le permite a la acción expresar cierta actitud, o creencia, o valor, o emoción, o cualquier cosa por el estilo. La expresividad, no la utilidad, sería lo que fluiría hacia atrás. Lo que fluye hacia atrás, hacia la acción, vehiculado por la conexión simbólica sería la posibilidad de expresar cierta particular actitud, o creencia, o valor, o emoción, etc. Expresar eso tendría una gran utilidad para la persona, razón por la cual realizaría la acción simbólica.[34]

Podría parecer que no hay mucha diferencia entre esos dos modos de estructurar o comprender por qué se elige realizar una acción simbólica. Cada uno de ellos proporcionará una explicación distinta de por qué *no* se realiza una acción simbólica. Por lo pronto, en la medida en que se imputa retroactivamente, unciéndola a la conexión simbólica, utilidad a la acción, nos enfrentamos a un dilema. Presumiblemente, la conexión simbolizadora se mantiene siempre, de manera que la acción de lavarse las manos simboliza siempre, por ejemplo, desprenderse de culpa. Puesto que es de presumir que esta situación simbolizada, verse libre de culpa, tiene siempre mucha utilidad, si la utilidad se imputa retroactivamente, ¿por qué no tiene siempre la acción de lavarse las manos una utilidad máxima, viéndose la persona impelida a hacerlo constantemente? (Aparentemente, esto es lo que les pasa a quienes se lavan compulsivamente las manos, pero no a todos, y no con todas las acciones realizadas por su valor simbólico.) La teoría de la expresividad dice que la posibilidad de expresar alguna actitud respecto a sentirse libre de culpa está siempre presente, como resultado de la sempiterna conexión simbólica, pero que la utilidad de expresarla de ese modo varía de contexto a contexto, según lo reciente o relevantemente que uno la

haya expresado, según cuales sean los demás deseos o necesidades de uno, etc. La utilidad de expresar esa actitud o emoción compite con otras utilidades. La teoría de la imputación de utilidad describirá eso de otro modo. La utilidad absoluta o relativa, para la persona, de la situación simbolizada puede fluctuar; la utilidad de estar libre de culpa puede llegar a decrecer realmente si la persona ha dado recientemente pasos para aliviar su culpa —aquí hay (por el momento) menos que hacer—. O la utilidad de estar libre de culpa puede permanecer constante mientras que la de otros bienes competitivos, como comer, crece momentáneamente hasta rebasar la utilidad de eliminar la culpa. Las dos estructuras de intelección de la expresividad simbólica concederán margen de fluctuación a la utilidad —un margen distinto—. Lo que me importa resaltar aquí es la *importancia* de este significado simbólico, independientemente de cómo haya que entender exactamente su vertebración.

Cuando, de acuerdo con su significado simbólico, se imputa utilidad a una acción o a un resultado —es decir, cuando se hace una ecuación entre la utilidad de una acción o un resultado y la utilidad de lo que simbólicamente significan— estamos prontos a pensar que esto es irracional. Cuando ese significado simbólico entraña deseos y temores infantiles reprimidos, o deseos y temores presentes pero inconscientes, puede traer consigo una conducta condenada a la frustración, a la insatisfacción o al tormento. Mas ¿no podrían por ventura los significados simbólicos basados en deseos inconscientes añadir reverberaciones gratificantes a los bienes conscientemente deseados? En cualquier caso, no todos los significados simbólicos arraigarán en material freudiano. Muchos de ellos, no obstante, parecerán extraños a quienes estén fuera de esa red de significados. No hay que olvidar cuán calamitosas consecuencias alguna gente está dispuesta a arrostrar con tal de evitar el «desprestigio», ni los riesgos mortales a los que a veces se enfrentan cuando se baten en duelos para «salvar el honor» o se enredan en hazañas para «probar su hombría». Pero no deberíamos apresurarnos a concluir que sería mejor vivir sin significados simbólicos de ningún tipo, o sin imputar nunca utilidades simbólicas de acuerdo con los significados simbólicos.

Los principios éticos codifican el modo de comportarse con los demás de una manera adecuada a su valor y a nuestro sentido de compañerismo para con ellos. Mantener y secundar principios éticos, además de servir a propósitos particulares, tiene un significado simbólico para nosotros. Tratar a la gente (y valorarla en general) con respeto y responsabilidad nos pone a nosotros «del lado de»

ese valor, aliándonos acaso con todos los que están de ese lado, y simboliza la urdimbre que con ellos formamos. (¿La simboliza exagerando lo que es esa urdimbre realmente, o acaso una conexión simbólica aceptada constituye una urdimbre real?) Kant sentía que, al actuar moralmente, una persona actúa como miembro del reino de los fines, como un legislador libre y racional. La acción moral no es la *causante* de que nos convirtamos en un miembro (permanente) de ese reino. Es lo que nosotros haríamos como tales miembros, es un paradigma de lo que se haría en tales circunstancias, y por lo mismo, simboliza el hacerlo en tales circunstancias. Las acciones morales se agrupan con otros posibles acontecimientos y acciones, y llegan a valer por ellos y a significarlos. De aquí que el ser ético adquiera una utilidad simbólica conmensurable con la utilidad que realmente tienen esas otras cosas por las que vale. (Eso depende, entonces, de que esas otras cosas tengan realmente utilidad para la persona —una contingencia en la que Kant estaría poco dispuesto a confiar—.) Hay una variedad de cosas que una acción ética podría significar simbólicamente para alguien: ser una criatura racional que se da a sí propia leyes; ser un miembro legislador de un reino de los fines; constituirse, en pie de igualdad, en fuente y en reconocedor de valor y de personalidad; ser una persona racional, desinteresada y no egoísta; ser humanitario; vivir de acuerdo con la naturaleza; responder a lo valioso; reconocer a alguien como criatura de Dios. La utilidad de esas grandes cosas, simbólicamente representadas y ejemplificadas por la acción, viene a incorporarse a la utilidad (simbólica) de esa actividad. Así van formando parte esos significados simbólicos de las propias razones para actuar éticamente. Ser ético es una de las maneras más efectivas de simbolizar (una conexión) con aquello que nos resulta más valioso.

Una gran parte de la riqueza de nuestras vidas consiste en significados simbólicos y en su expresión, en los significados simbólicos que nuestra cultura atribuye a las cosas, o en los que nosotros mismos les otorgamos.* En cualquier caso, no resulta claro qué podría querer decir vivir sin significados simbólicos, que ninguna parte

* Obsérvese que los significados simbólicos podrían no ser todos significados buenos. Lo mismo ocurre con los deseos y las preferencias. El punto importante es que una teoría de la racionalidad está obligada a no *excluir* los significados simbólicos. Sin embargo, éstos no garantizan un contenido bueno o deseable. Lo que se necesitaría entonces es desarrollar una teoría de qué significados simbólicos y de qué preferencias y deseos resultarían admisibles, utilizando ese resultado para restringir los significados y deseos particulares que podrían tener cabida en la teoría más formal de la racionalidad.

o medida de nuestros deseos dependiera de tales significados. ¿Qué tendríamos entonces que desear? ¿Simplemente comodidad material, seguridad física y placer sensual? ¿Y no dependería parcialmente la intensidad con que los deseamos del modo en que simbolizan amor y cuidado maternos? ¿Simplemente riqueza y poder? ¿Y no vendría parte de la intensidad con que los deseamos del modo en que éstos podrían simbolizar la desvinculación respecto de la dependencia infantil, o el éxito en la competición con uno de los padres? ¿En ningún caso de los significados simbólicos que el poder y la riqueza puedan traer consigo? ¿Cuentan simplemente los reforzadores innatos, incondicionados, que la evolución ha destilado e inculcado en nosotros, contando todo lo demás sólo en la medida en que resulta un medio adecuado para ellos? Estos reforzadores habrían servido para que nuestros ancestros fueran progenitores más eficaces y más eficaces protectores de genes emparentados. ¿Deberíamos elegir esto como nuestro único propósito? Y si lo tuviéramos en alta estima, ¿no podría también resultarnos estimable cualquier cosa que simbolizara el ser un progenitor real? «No, no si eso entrara en conflicto con ser un progenitor real, y en cualquier caso, a uno debería resultarle sólo estimable la producción real y la protección de la progenie propia y allegada, así como los medios efectivos que la evolución ha filtrado para eso, a saber, los reforzadores incondicionales y los medios de que *éstos* se sirven.» (Obsérvese, no obstante, que el tener inculcados deseos que sirven para maximizar la adaptación inclusiva no significa que la evolución nos haya inculcado también el deseo de maximizar la adaptación inclusiva. Presumo que los varones no están haciendo ahora mismo colas larguísimas a la puerta de las clínicas de inseminación artificial para convertirse en donantes de esperma, aunque esto serviría para incrementar su adaptación inclusiva.) ¿Pero por qué el hecho de que llevar realmente a algo sea mucho mejor que simbolizarlo habría de implicar que la simbolización no cuenta para nada? «Porque ésta es la línea última, lo que ocurre realmente; todo lo demás es charlatanería.» Pero, ¿por qué esta línea última es mejor que todas las demás?

En cualquier caso, si somos criaturas simbólicas —y la antropología es testigo de la naturaleza universal de ese rasgo—, entonces, presumiblemente, la evolución nos hizo así. Por consecuencia, los atractivos placeres de la simbolización, así como las satisfacciones simbólicas, están tan sólidamente fundados como el resto de reforzadores innatos. Quizá la capacidad de simbolización sirvió para robustecer otros deseos, o para mantenerlos a través de períodos de falta de refuerzos por parte de sus objetos reales. Sin embargo, cual-

quiera que sea la explicación evolucionaria, esta capacidad, como otras capacidades cognitivas, no está atascada en su función adaptativa original. Puede emplearse de otros modos valiosos, lo mismo que las capacidades matemáticas pueden emplearse para explorar la teoría abstracta de los números y las teorías del infinito, aunque no fue ésa la función por la que fueron seleccionadas. Una vez que existe la capacidad para la utilidad simbólica, ésta puede permitirnos, por ejemplo, lograr en algún sentido —esto es, simbólicamente— lo que es causal o conceptualmente imposible, extrayendo así utilidad de ello, y puede permitirnos también separar los rasgos buenos y los rasgos malos con que realmente van amalgamados, extrayendo de esa amalgama sólo los primeros a través de algo que los simboliza sólo a ellos.

Eso no quita para negar los peligros potenciales de los significados simbólicos y las utilidades simbólicas. Los conflictos pueden llegar a implicar rápidamente significados simbólicos que, exagerando crecientemente la importancia de los problemas, induzcan a la violencia. Los peligros que hay que evitar especialmente tienen que ver con situaciones en las que las consecuencias causales de una acción son extremadamente negativas, pero el significado simbólico de la misma es tan grande que, aun así, la acción se acaba de todas formas realizando. (Recuérdense los ejemplos de la acción compulsiva de lavarse las manos y de la prohibición de la droga.) Una persona racional buscaría una alternativa simbólica (casi) tan satisfactoria pero que no tuviera consecuencias reales tan calamitosas. (Esto no implica, sin embargo, que los significados simbólicos tengan que andar siempre subordinados a, y venir lexicográficamente después de, los resultados causalmente producidos.) A veces, se pensará que una conexión simbólica es mejor que una conexión causal. Si un resultado —dañar por venganza a otro, pongamos por caso— se ve como deseable, pero también como dañino, podría ser mejor para la persona vengarse simbólicamente que infligir un daño real.[35] Estaría bien descubrir un criterio general estructural acerca de los tipos de vínculos que forjan los significados simbólicos, un criterio que permitiera distinguir entre los significados simbólicos buenos y los malos, pero quizá deberíamos limitarnos a estar atentos a ciertos tipos de situaciones —el conflicto es una de ellas— para aislar y excluir significados simbólicos particulares. Podría ser de ayuda el hecho de que muchos significados simbólicos indeseables no estén en equilibrio una vez conocidas sus causas; si supiéramos el origen de sus significados, o el papel que desempeñan en nuestras acciones presentes, no querríamos actuar de acuerdo con

ellos.[36] Algunos significados simbólicos pasan esos tests (por ejemplo, el significado simbólico de un gesto romántico dirigido a una persona que ustedes aman). Quizá lo crucial sea mantener la alerta para saber cuándo son simbólicas las conexiones, guardando para éstas una pista aparte para no tratarlas (inadvertidamente) como conexiones causalmente reales. Esto nos ayudaría a entender los varios significados simbólicos freudianos, los cuales, cuando entran en la deliberación consciente como tales, pierden su poder y su impacto (si son suficientemente «elaborados»).[37]

El significado simbólico es también un componente de las decisiones éticas particulares. Se ha sostenido que el significado simbólico de los esfuerzos para salvar a una persona que está corriendo peligro —un minero atrapado, por ejemplo—, o el de negarse a hacerlos, afecta a nuestra decisión de asignar recursos alternativamente o bien a medidas de salvamento, o bien a medidas de prevención de accidentes. (Este asunto ha sido conceptuado como un asunto de oposición entre «vidas reales y vidas estadísticas».)[38] Se ha sostenido también que el significado simbólico de alimentar a alguien, de darle sustento, entra en la discusión de las formas en que puede ponerse permisiblemente fin a la vida de personas calamitosamente enfermas —desconectando el respirador artificial, por ejemplo, pero no interrumpiendo su nutrición para dejarles morir de inanición—.[39] La filosofía defendida en *Anarquía, Estado y Utopía* ignoraba la importancia que reviste para nosotros la afirmación y la expresión común y pública de nuestros vínculos e interés social, y por eso (he escrito que) resulta inadecuada.[40]

Vivimos en un mundo pletórico de símbolos, en parte culturales y en parte de propia creación, y merced a ello, escapamos de o aplazamos los límites de nuestras situaciones, y no simplemente fantaseando, sino actuando y sirviéndonos de los significados de las acciones. A las acciones y a los acontecimientos les imputamos utilidades coordinadas con lo que esas acciones y esos acontecimientos simbolizan, y aspiramos a realizarlos (o a evitarlos) según nuestras aspiraciones a aquello por lo que ellos valen.[41] Así, pues, se necesita una teoría más amplia de la decisión para poder incorporar esas conexiones simbólicas y para detallar la nueva vertebración a que éstas invitan.

Entre los científicos sociales, son los antropólogos quienes mayor atención han prestado a los significados simbólicos de las acciones, de los rituales y de las formas y prácticas culturales, así como a su importancia en el discurrir de la vida de un grupo.[42] Su trabajo tiene tal grado de elaboración, que resulta un poco embarazo-

so introducir una noción relativamente ruda e indistinta de significado simbólico. Con todo, la noción se presta a varios usos, unos usos que no estarían bien servidos por discusiones matizadas y tejidas que no encajarían fácilmente con estructuras formales. Al incorporar el significado simbólico de una acción, su utilidad simbólica, a la teoría (normativa) de la decisión, podríamos vincular más estrechamente las teorías de la elección racional a las preocupaciones de la antropología. Estos vínculos pueden establecerse en dos direcciones. La primera, la dirección ascendente, explica las pautas y las estructuras sociales en términos de elección individual que incorpora la utilidad simbólica. Ésta, la dirección metodológicamente individualista y reductora, no es la que voy a proponer aquí.[43] La segunda, la dirección descendente, explica el modo en que las pautas y los significados sociales perfilados por los antropólogos tienen un impacto en las acciones y en la conducta de los individuos, es decir, a través de las decisiones de éstos que confieren cierto peso a la utilidad simbólica. (Algunos antropólogos, por una cuestión de orgullo profesional, parecen desentenderse del hecho de que los significados culturales por ellos perfilados están mediados por la conducta individual.)

¿Cómo funciona la utilidad de una acción (o de un resultado)? ¿Cuál es la naturaleza de la conexión o de la cadena de conexiones simbólicas? ¿Y de qué modo fluye la utilidad, o la posibilidad de expresividad, a través de esa cadena, desde las situaciones objeto de simbolización hasta las acciones (o los resultados) que desempeñan la tarea simbolizadora? Obsérvese, por lo pronto, que el significado simbólico va más allá del resultado conseguido por la adopción de principios, a saber, que algunas acciones valgan por otras. En la adopción de principios, una acción vale por otras cosas del mismo tipo —otras acciones—, o para un grupo entero de ellas, mientras que el significado simbólico puede conectar una acción con cosas distintas de (un grupo de) acciones, por ejemplo, con ser un cierto tipo de persona, con la realización de un cierto estado de cosas.

Nelson Goodman ha propuesto algunas categorías útiles y sugerentes.[44] De acuerdo con Goodman, *A denota B* cuando *A* se refiere a *B*; *A ejemplifica P* cuando *A* se refiere a *P* y *A* es un caso de *P*, esto es, es denotado por *P* (ya literal, ya metafóricamente); *A expresa P* cuando *A* se refiere a *P* y *A* tiene la propiedad *P* figurativa o metafóricamente (de modo que *P* denota figurativamente a *A*), y, al ejemplificar *P*, *A* funciona como un símbolo estético. Esas relaciones pueden encadenarse. *A alude a B* cuando *A* denota algún *C* y este *C* ejemplifica a *B*, o cuando *A* ejemplifica a algún *C* y este *C*

denota a *B*. Aún son posibles cadenas más largas,[45] algunos de cuyos vínculos pueden ser figurativos o metafóricos. Esas cadenas, y otras, pueden conectar una acción con situaciones o condiciones ulteriores más amplias, aquellas que puede representar simbólicamente y a las que puede aludir (etc.), y la utilidad de esas situaciones más amplias proporciona entonces a la acción misma una *utilidad simbólica* que entra en las decisiones acerca de esa acción. No es necesario que esas cadenas sean muy largas: cuando *A* está en la extensión literal de un término *P*, y *B* está en la extensión metafórica de ese término, *A* podría tener a *B* como parte de su significado simbólico. A veces, una acción puede significar simbólicamente algo por el hecho de ser nuestra más paradigmática realización de ese algo, lo mejor que podemos hacer.[46]

¿De qué particular modo viene determinada la utilidad simbólica (o expresividad) de una acción por la utilidad de esa situación más amplia, a la que la cadena conecta la acción, y por la naturaleza de la cadena misma? ¿Acaso las cadenas más cortas transmiten más utilidad/expresividad de la situación amplia a la acción misma? ¿Se pierde tanta más utilidad/expresividad cuantos más eslabones haya? Diferentes tipos de eslabones, ¿transmiten diferentes proporciones de (o posibilidades de expresar) la utilidad de la situación más amplia? (Parto del supuesto de que la utilidad simbólica de una acción no puede ser mayor que la utilidad de la situación más amplia a la que está conectada por la cadena, y de que puede ser menor.) ¿No todas las conexiones simbólicas consiguen inducir la retroimputación de utilidad, y en el caso de que lo consigan, qué determina cuáles? Esas cuestiones se plantean siempre en situaciones de elección en condiciones de certidumbre; otro tipo de cuestiones surgen en condiciones de incertidumbre. ¿Hay algún descuento probabilístico a lo largo de algunas cadenas particulares? Algunos tipos de situaciones más amplias, aun en el supuesto de que no resultaran con certeza, ¿consiguen transmitir retroactivamente su plena utilidad a las acciones de que ellas mismas habrán de resultar? Por lo demás, ni que decir tiene que el mismo hecho de que una acción entrañe particulares riesgos e incertidumbres puede conferirle una particular utilidad y un particular significado simbólico, que acaso tenga que ver con ser una persona osada y valiente, o hasta temeraria. A veces, sin embargo, la presencia de probabilidades, no de certidumbre, puede eliminar completamente el significado simbólico. *No* es verdad que una probabilidad de un medio, o de un décimo, de realización de cierto objetivo tenga siempre un medio o un décimo de la utilidad de ese objetivo —no necesariamente simboliza el

objetivo, ni siquiera parcialmente—. He aquí otra razón para tratar a las utilidades simbólicas como un componente separado en una teoría de la decisión, en vez de incorporarlas sin más a las presentes teorías (causales y probatorias) de la decisión. Pues estas utilidades simbólicas no obedecen a una fórmula de valor esperado. Podríamos tratar de entender y explicar *determinadas* desviaciones observadas respecto de una fórmula de valor esperado y respecto de los axiomas conexos de la teoría de la decisión atribuyendo esas desviaciones a las utilidades simbólicas. Pienso ahora en la paradoja de Allais, en el efecto de certidumbre, en determinadas desviaciones del principio *sure thing* o de seguridad de Savage, etc. La *certidumbre* misma tiene para nosotros una utilidad simbólica. La diferencia que media entre las probabilidades 0,9 y 1,0 es mayor que la que media entre 0,8 y 0,9, aunque esa diferencia entre diferencias desaparece cuando todas están incrustadas en idénticos y más amplios juegos probabilistas de apuesta —el que desaparezca es lo que distingue como simbólica a la diferencia—.* Está por desarrollar una

* ¿O acaso se trata simplemente de que algunos números están más distinguidos y de que la utilidad se ve afectada por esa distinción? Una inflación de dos dígitos tiene el significado simbólico de una inflación descontrolada, de manera que causa mayor preocupación un incremento de la misma que la haga pasar del 9 al 10 por ciento que un incremento del 16 al 17 por ciento; si contáramos con base once, la línea (simbólica) estaría trazada en otro sitio. En *Anarquía, Estado y Utopía* me referí al significado de *eliminar* un problema completamente, en el sentido de que hay mayor diferencia en reducir el número de casos de un mal de uno a cero que en reducirlo de dos a uno. Me refería a eso como a la marca de un ideólogo (pág. 266); es mejor verla como la marca de un significado simbólico.

Obsérvese que el efecto de certidumbre, cuando acontece, requiere que la utilidad sea medida con un procedimiento ligeramente distinto del habitual. En el procedimiento habitual, se asignan a dos resultados, x y z, números de utilidad ordenados de acuerdo con la preferencia entre esos resultados, y la utilidad de cualquier otra cosa, y, se determina de acuerdo con la condición arquimédea. Esa condición dice que, cuando se prefiere x a y, e y se prefiere a z, entonces hay una única probabilidad p (entre 0 y 1, ambos exclusive) tal que la persona es indiferente entre y con certeza y una opción consistente en una probabilidad p de x y una probabilidad $(1-p)$ de z. Cuando la persona satisface plenamente todas las condiciones de Von Neumann-Morgenstern, no hay ningún problema; pero cuando se da el efecto de certidumbre, se asignará a esa opción intermedia, y, una utilidad equivocada. Un procedimiento mejor podría consistir en medir la utilidad prescindiendo de resultados ciertos, e incrustando todo lo anterior dentro de mezclas canónicas de probabilidad, por ejemplo, en la probabilidad 1/2. Se le pediría entonces a la persona que hallara el valor de la probabilidad p, en la que ella misma fuera indiferente entre, por un lado, no conseguir nada con probabilidad 1/2 y conseguir y con una probabilidad 1/2, y por otro lado, no conseguir nada con probabilidad 1/2 y 1/2 de probabilidad de conseguir x con probabilidad p y z con probabilidad $1-p$. Así conseguiríamos controlar el efecto de certidumbre. Evidentemente, tal procedimiento podría funcionar sólo

teoría detallada de la utilidad simbólica. Lo que podemos hacer por el momento es acotar un espacio para ella dentro de la estructura de una teoría más general de la decisión, un espacio sobre el que tendré ocasión de decir más en el próximo capítulo.

Mecanismos teleológicos

Los principios les ayudan a ustedes a descubrir la verdad transmitiendo apoyo probatorio o probabilidad de unos casos a otros. Los principios les ayudan también a ustedes a vencer la tentación transmitiendo utilidad de algunas acciones a otras. Los principios son mecanismos de transmisión de probabilidades y de utilidades.*

Los principios tienen varias funciones y varios efectos: intelectuales, intrapersonales, personales e interpersonales. Eso no quiere decir que tengan todos esos efectos en cualquier posible situación. Un mecanismo regulador de temperatura funciona sólo dentro de un determinado espectro de temperaturas; más allá de ese espectro, no será capaz de recuperar la temperatura de nuevo, y según el ma-

---

si no fuera sensible a la particular probabilidad —1/2, en este caso— dentro de la mezcla canónica de probabilidades. Tendría que darse el caso de que los mismos resultados se dieran dentro de una gran variedad de probabilidades en la mezcla canónica, quizá con todos los que estuvieran en el intervalo ε de 0 o 1.

* ¿Deben todos los principios limitarse a transmitir unas u otras, o pueden algunos principios transmitir ambas? ¿Deberíamos permitirnos la especulación de que hay *una* cosa que todos los principios transmiten, a saber: $p_i \times u^i$, probabilidad *y* utilidad? La teoría de la decisión carece de un concepto para designar esa adición ponderada, $p_i \times u^i$, aun a pesar de que sus componentes viajan frecuentemente *como una unidad.* En realidad, las teorías formales tienen que instituir procedimientos muy particulares para poder deslindarlos, procedimientos que con frecuencia parten del supuesto de que ya han sido deslindados con éxito en determinados casos para luego usar mecanismos que permiten extender esos casos a situaciones genéricas. Podría resultar muy instructivo tratar las probabilidades y las utilidades como parte de una cantidad integrada —llamémosle importancia—, no separando demasiado pronto sus componentes e investigando qué condiciones satisface esa cantidad integrada. (¿Hay, empero, una asimetría inicial entre los componentes, en la medida en que la importancia puede ser incrustada en las mezclas de *probabilidad*? ¿Necesitamos investigar las correspondientes posibilidades de las mezclas de utilidad, que pueden exagerar o disminuir las importancias de los componentes? ¿Podría acaso incluirse inicialmente un factor temporal en la combinación, para luego abstraerse de él en tanto que componente? Las preferencias temporales, ¿son asunto de la utilidad o de la probabilidad, o constituye la distancia temporal misma una disminución en la importancia? La extensión de una utilidad en el tiempo —no su desplazamiento por el tiempo—, ¿exagera su importancia?)

terial de que esté hecho, él mismo puede fundirse o congelarse. ¿Por qué no nos dotó la evolución de mejores mecanismos reguladores de la temperatura corporal? Dado lo pequeño de la probabilidad de que ocurran casos extremos como los mencionados, eso resultaría demasiado costoso en términos de energía y de sacrificio alternativo de otras funciones. Un mecanismo puede cumplir su función bastante bien, lo suficientemente bien, aun si no funciona en algunas de las situaciones que podrían presentársele. Lo mismo ocurre con los principios.

Para poder justificar un principio, ustedes definen una función y muestran que el principio cumple efectivamente esa función, y que lo hace más eficazmente que otros, dados los costes, restricciones, etc. También podemos preguntarnos por la deseabilidad de esa función. ¿Por qué *algo* debería cumplirla? Habrá que hacer una justificación y mostrar (o suponer) que la función es deseable y no se interfiere en otras funciones más deseables. Plenamente definida, una justificación de un principio *P* es una estructura de la teoría de la decisión, una estructura en la que el principio *P* ocupa el lugar de una acción, una acción que compite con determinadas alternativas y que tiene determinadas probabilidades de lograr determinados fines, los cuales resultarán deseables en determinadas medidas, etc. (Nuestra anterior discusión sobre los factores que habrían de ser considerados por una teoría de la elección óptima de principios encajaría en este marco teleológico de la teoría de la decisión.)

Puede diseñarse un principio para lidiar con determinadas situaciones, o para protegerse de peligros *particulares* tales como ceder a tentaciones del momento, favorecer los intereses propios, creer lo que uno quisiera que fuera verdad, etc. De aquí que quien no quiera enfrentarse a esos peligros podría no tener ninguna necesidad de *esos* principios. También podría haber otros mecanismos, distintos de los principios, para lidiar con estos peligros. (¿Podría una persona evitar actuar con favoritismo hacia sus propios intereses no sólo valiéndose de principios, sino mediante la interacción empática con otros, proyectándose plena e imaginativamente en las circunstancias de los demás?)

Podríamos preguntarnos si el mismo mecanismo de los principios generales tiene sus *propios* sesgos y defectos. Plantearnos las cosas en términos de la teoría de la decisión nos permite ver los principios en tanto que mecanismos (a los que se les supone) con ciertos efectos —sus funciones—, y por lo mismo, nos permite no sólo comparar algunos principios con otros, sino también comparar el mecanismo de los principios con otros mecanismos. Algunos fines po-

drían ser de imposible o harto difícil realización mediante principios, cuando, en cambio, otros medios podrían conseguir *esos* fines más fácilmente.

Si un fin importante es vivir en comunidad sin un conflicto tan intenso que desbiele y destruya instituciones sociales valiosas, entonces, cuando las partes en liza afirman principios incompatibles, quizá no haya ninguna forma de resolver ese conflicto haciendo coincidir a las partes en un tercer principio, y no digamos en alguno de los principios originales. Lo que acaso se necesite es algún compromiso, pero precisamente los compromisos es lo que se supone que los principios no pueden lograr. De aquí que un dirigente de una institución o de un país pueda limitarse a tratar de mantener las cosas en funcionamiento, maniobrar de algún modo para contener la furia del pueblo de manera que la vida institucional pueda seguir su curso. Es verdad que puede haber un principio que recomiende hacer eso, un principio que pueda aplicarse a cualquier situación de serio conflicto entre principios que amenace con tornar disfuncionales instituciones valiosas. El contenido particular del compromiso, sin embargo, puede estar simplemente determinado por aquello a lo que las fuerzas en liza, dada la correlación entre ellas, pueden conseguir adaptarse. No es necesario que tal compromiso esté determinado por el principio, en el sentido de que sus detalles vayan a constituir un precedente para otras situaciones similares.

Eso no significa recomendar que los dirigentes políticos e institucionales carezcan de principios. Quizá deban tener principios en sus decisiones y en sus acciones, salvo en aquellas raras situaciones en las que entra en escena el principio antes mencionado, de acuerdo con el cual hay que entrar en un compromiso (al margen de principios). (Sin embargo, si atendemos a la estructura del gobierno de los Estados Unidos, parece haber una división diferente: algunos tipos de decisiones —las tomadas por el poder judicial— no pueden proceder sin principios, mientras que los detalles de otras decisiones —las del ejecutivo y las del legislativo— se abandonan generalmente al juego de varias fuerzas, con cierta vigilancia por parte del judicial para garantizar que no serán violados determinados principios generales.) Lo único que deseo subrayar aquí es que el mecanismo teleológico de los principios podría no resultar adecuado para todos y cada uno de los propósitos.

¿Significa eso, entonces, que debemos emplear algún principio para decidir cuándo hay que invocar un principio? ¿Y qué es lo que hace que *esta* situación sea una situación guiada por principios en vez de por alguna otra cosa? La concepción, según la cual cualquier

cosa debe decidirse de acuerdo con un principio podría felizmente proponer eso también como un principio. Mas la concepción, según la cual algunas cosas, pero no otras, deberían decidirse de acuerdo con principios, ¿es ella misma una concepción de principios —y está obligada a serlo—? ¿Hay alguna *presunción* favorable a decidir las cosas de acuerdo con principios que pueda ser cancelada por razones específicas en favor de algún otro mecanismo en esa situación? Si no, la decisión de algún caso particular fundándose en principios, ¿no plantearía la cuestión de por qué se apeló *entonces* a precisamente un *principio*? ¿Podría haberse hecho esto para sesgar la particular elección realizada evitando el resultado al que hubiera llegado algún otro mecanismo? ¿O acaso no hay presunción de ningún tipo, sino simplemente una noción de teoría de la decisión respecto a cuándo resultan adecuados los principios, una noción que usa ella misma algún principio de la teoría de la decisión y presupone que, al menos en este caso —cuando se trata de decidir cuándo resultan apropiados los principios—, es apropiado usar algún principio, un principio de la teoría de la decisión?

Otra razón para pensar que los principios de acción tienen una función teleológica es la siguiente. Un caso real, por ejemplo el de la Alemania nazi, puede socavar o refutar de punta a rabo un principio *P*, según el cual todo estaría permitido. Mas, ¿por qué no bastaría el ejemplo hipotético? ¿Se podría haber dicho en 1911: el principio *P* permitiría, o en determinadas circunstancias incluso exigiría, (algo parecido a) la Alemania nazi; por lo tanto, *P* es falso, inaceptable, malo?

No obstante, si se supone que los principios se limitan a cubrir los casos que ocurrirán, que ocurrirían o que podrían ocurrir (es decir, a cubrirlos antefácticamente), y si se cree que un caso es imposible (que la situación, las motivaciones, o cualquier cosa por el estilo, que llevarían a él no surgirán, o no tendrán éxito), podría entonces no ser considerado un contraejemplo *relevante* de ese o de cualquier otro principio. Una vez, empero, se ha descubierto que la naturaleza humana *puede* hacer eso —puesto que lo *hizo*—, entonces el principio *P* que lo permite queda refutado.

Las consecuencias de que la gente acepte y actúe de acuerdo con un principio pueden desacreditar ese principio. «Actuaron de acuerdo con el principio *P*, y vean la tremebunda situación a que eso ha dado lugar.» Alguien podría reponer que se llevó demasiado lejos el principio, o que fue secundado de un modo incorrecto, que el principio mismo no *requería* lo que se hizo en su nombre. Sin embargo, el principio *P* ha quedado desacreditado. Cuando alguien se indig-

na con las consecuencias que ha acarreado secundar *P*, resulta difícil decirle: «Secundemos un nuevo *P*, pero esta vez correctamente». ¿Por qué? ¿Acaso porque *P* se *plegó él mismo* tan fácilmente a este modo de actuar, aun no requiriéndolo él mismo? Esto es lo que lleva a la gente que secunda *P* a actuar de este modo cuando la gente es realmente como es.[47] Si un principio es un mecanismo destinado a tener determinados efectos, es un mecanismo destinado a tener dichos efectos *cuando es secundado*; de manera que lo que ocurre realmente cuando es secundado, no lo que el principio mismo *dice*, es lo relevante a la hora de estimar el principio en tanto que mecanismo teleológico.

Mas, ¿no son también los principios verdades básicas, verdades que nos ayudan a comprender por la vía de cobijar casos (*à la* Hempel), explicando así por qué estos últimos se dan? También aquí podríamos considerar a los principios como mecanismos con la función *epistemológica* de producir comprensión, de manera que también aquí habría que preguntarse (en términos de teoría de la decisión) si hay otras rutas abiertas a la comprensión, si esas otras rutas resultan más adecuadas para algunos contextos u objetos, etc.

Mas ¿no podrían ser los principios lo que hace verdaderas a las verdades particulares, lo que les da origen, en cuyo caso la primacía de los principios sería *ontológica*? Si esto no se limita a repetir simplemente la función epistemológica —comprendemos del mejor modo las verdades particulares merced a los principios—, y si el «dar origen a» no es una relación temporal y el «hacer verdaderas» no es una relación causal, entonces no resulta exactamente claro lo que sostiene la tesis ontológica. En cualquier caso, lo que no podemos negar es que la formulación de principios (para las matemáticas, para los fenómenos naturales, para la psicología) puede dar coherencia a esos fenómenos y profundidad a nuestra comprensión, sea esa relación entre fenómenos y principios una relación ontológica, una relación epistemológica, o alguna relación de tipo mixto. De aquí que los principios cumplan otra función intelectual además de la ya estudiada (la transmisión de apoyo probatorio y de probabilidad), a saber: profundizar, unificar y hacer explícita nuestra comprensión de aquello de que se ocupan los principios. (Esa función generará relaciones más robustas de apoyo probatorio y de probabilidad. ¿Es posible que éstas *constituyan* a, más que resulten del incremento de comprensión?) Así, la formulación de principios morales podría hacer más profunda nuestra comprensión de la acción moral o de los hechos y fenómenos morales. Ello es, sin embargo, que los principios morales tendrían un estatus indistinguible del de los físicos o

psicológicos que describen los fenómenos pero que no aportan razones para actuar *de acuerdo con ellos*. Podría decirse que aunque los principios morales correctos son verdaderos —en el sentido de que *deben* ser secundados—, el único modo de realizarlos, es decir, de ser verdaderos en nuestra conducta real, es tratar de secundarlos, de actuar *de acuerdo con ellos*. Es ésta una afirmación empírica, una afirmación que requeriría evidencia probatoria. Quizás los principios que somos capaces de formular y secundar andan tan alejados de lo que los principios morales correctos —verdades morales más complejas— requerirían, que mejor sería conformarnos con estos últimos siguiendo rutas distintas de las que tratan de actuar de acuerdo con principios. Se trata, después de todo, de una cuestión empírica. En cualquier caso, esto hace que actuar de acuerdo con un principio sea, una vez más, un mecanismo teleológico.

El término *principios* se usa a menudo para referirse a algo más profundo y más general que las reglas. Esos principios son los perfiles generales dentro de los cuales hallan un lugar los detalles. En las negociaciones de acuerdos, esos perfiles generales pueden convertirse en líneas orientativas; las partes comienzan acordando principios que gobernarán el tratado de paz entre países, la fusión entre corporaciones, el nuevo tipo de escuela que hay que establecer; y luego seguirán los detalles. Acordar ese marco general de principios puede ahorrar tiempo después si las elecciones acerca de los detalles están obligadas a optar entre alternativas que caigan bajo esos principios generales, de manera que no todas las posibilidades queden abiertas a la discusión y al debate, ni quede todo asunto pendiente de reaparecer. Además, los acuerdos sobre principios pueden invocarse siempre para dirimir disputas. Pero los principios en general no están confinados a la guía orientativa de la acción. Los títulos de los libros de texto anuncian que presentarán los principios de la teoría económica, de la física o de la psicología, el marco de los enunciados generales en el que habrán de caber subordinadamente los detalles de la materia. Quizás haya aquí también cierta guía orientativa: los principios presentados orientarán la comprensión por parte del lector de los detalles de la materia estudiada.

Si nuestras capacidades cognitivas estuvieran limitadas de determinadas maneras, podría haber una diferencia entre los principios que podrían orientar nuestra comprensión de una materia y los enunciados generales verdaderos (o leyes) que subyacen a los hechos de la materia en cuestión. Estas últimas generalizaciones podrían llegar a resultarnos demasiado complicadas y difíciles de comprender, no conseguiríamos derivar consecuencias de las mismas

y llenarlas con los detalles. ¿Cuáles serían entonces los principios de (por ejemplo) la física; las generalizaciones razonablemente exactas que podríamos entender y bajo las que podríamos subsumir detalles, o los enunciados precisamente verdaderos y omniabarcantes que no podríamos comprender y manejar?

La tradición kantiana tiende a sostener que la función de los principios es orientar la deliberación y la acción de criaturas autoconscientes y reflexivas; de aquí que los principios tengan una función teorética y una función práctica. Somos criaturas que no actúan automáticamente, sin orientación. Podemos imaginarnos en posesión de una orientación automática (¿nos resultarían entonces los principios completamente ociosos?), o, más pertinente aún, actuando de tal modo que prescindiéramos de orientación, por ejemplo, al azar. (¿Bastaría el actuar completamente al azar para liberarnos del dominio de la causalidad, la función que Kant reservó a los principios?) ¿No revela esto que el propósito de los principios es orientarnos y guiarnos hacia algo, lo que sea, que no conseguiríamos alcanzar actuando al azar? ¿Y no reduce eso los principios a mecanismos teleológicos? Kant, sin embargo, sostendría también que los principios son una expresión de nuestra naturaleza racional, constitutiva de racionalidad. Pensar o actuar racionalmente consiste precisamente en conformarse a (ciertos tipos de) principios. De aquí que sea un error limitarse a atender a las *funciones* extrínsecas cumplidas por los principios. Si los principios son algo que sólo un agente racional puede formular y emplear, y si ser racionales es algo que tenemos en estima, entonces secundar principios puede simbolizar y expresar nuestra racionalidad. De manera que los principios podrían revestir una gran utilidad para nosotros no sólo a causa de aquello propiciado y conseguible por su uso, sino a causa también de lo que tal uso simboliza y expresa. En esa medida, los principios no serían meros mecanismos teleológicos. Pero aún subsistiría la cuestión de por qué valoramos tanto nuestra naturaleza racional y la actuación a partir de principios y razones que la expresan, si nuestra naturaleza racional no sirve a ningún propósito ulterior. ¿Por qué habría de detenerse aquí el gamo?

¿Por qué están los principios tan estrechamente vinculados a la racionalidad? ¿Y por qué valoramos la racionalidad? Decir de algo, una acción o una creencia, que es racional es lo mismo que evaluar las razones por las que se realiza o se alberga (y también el modo en que la persona toma en cuenta las razones en *contra* de realizarla o de albergarla). Si, por su naturaleza, las razones son generales, y si los principios capturan la noción de actuar *por* tales razones

generales (de modo que la persona se compromete a actuar así en otras circunstancias relevantemente similares), entonces, para actuar o pensar racionalmente, uno tiene que hacerlo de acuerdo con principios. Pero, ¿por qué deberíamos creer o actuar racionalmente? Una respuesta podría decir que *somos* racionales, que tenemos la capacidad para actuar racionalmente, y que valoramos lo que hacemos.[48] Mas, si tenemos que rebasar el simple autoelogio, ¿no deberíamos invocar las funciones a las que sirve el creer o el actuar racionalmente? ¿Y por qué deben ser generales las razones? Compáreselas con sus parientes más allegados no generales. Para explicar por qué deberíamos usar razones y no esas alternativas, debemos apelar de nuevo a las funciones de las razones.

Así, pues, en resolución, la cuestión pasa de ser una cuestión acerca de los principios a ser una cuestión acerca de la racionalidad. ¿Para qué sirven las razones? ¿Cuál es la función de la racionalidad? ¿Es la racionalidad misma enteramente teleológica, enteramente instrumental? Ésas son las cuestiones que motivan los capítulos que siguen.

Capítulo 2

EL VALOR DECISIONAL

Los economistas y los especialistas en estadística han desarrollado una elaborada teoría de la decisión racional, y han conseguido difundirla en las investigaciones teóricas y de política aplicada. Se trata de una teoría potente, matemáticamente precisa y manejable. Aun cuando su adecuación como teoría descriptiva de la conducta humana ha sido ampliamente cuestionada, vige como el punto de vista dominante acerca de las condiciones que una decisión racional debería satisfacer: es el punto de vista normativo dominante. Yo creo que esta teoría estándar de la decisión necesita ampliarse para incorporar explícitamente consideraciones acerca del significado simbólico de las acciones y algunos otros factores. El problema de Newcomb puede proporcionar una instructiva introducción a las insuficiencias de la teoría estándar. Me propongo formular aquí una teoría más amplia de la decisión que permita abordar y comprender adecuadamente este problema, para luego aplicar esa teoría ampliada a la iluminación del dilema del prisionero, un problema que da un incisivo perfil a las cuestiones de cooperación social racional y a la cuestión de si la coerción resulta (a veces) necesaria para mantener la cooperación, y que, por lo mismo, ha impulsado en los últimos años un buen número de elaboraciones formales de la teoría social.

El problema de Newcomb

El problema de Newcomb es muy conocido, y me limitaré a describirlo brevemente aquí.* Un ser, en cuyos poderes para predecir sus elecciones correctamente ustedes han depositado una gran confianza, va a predecir su elección en la situación que sigue. Hay dos

* El problema fue concebido por William Newcomb, un físico, me fue referido a mí por un amigo común y, con el permiso de Newcomb, lo presenté y discutí por vez primera en Robert Nozick, «Newcomb's Problem and Two Principles of Choice», en *Essays in Honor of C.G. Hempel*, comp. por N. Rescher y otros (Dordrecht: Reidl, 1969), págs. 114-116.

cajas, B1 y B2. La caja B1 contiene 1.000 dólares; la caja B2, o bien un millón de dólares, o bien nada. Ustedes pueden elegir entre dos acciones: (1) tomar lo que hay en las dos cajas; (2) tomar sólo lo que hay en la segunda caja. Además, ustedes saben, y el ser sabe que ustedes saben, y así sucesivamente, que si el ser predice que ustedes tomarán lo que hay en las dos cajas, no pondrá el millón de dólares en la segunda caja; si el ser predice que ustedes tomarán sólo lo que está en la segunda caja, pondrá el millón de dólares en la segunda caja. Primero el ser realiza su predicción; luego pone o no el millón de dólares en la segunda caja, de acuerdo con su predicción; luego realizan ustedes su elección.

El problema no consiste sólo en decidir qué hacer, sino en entender exactamente qué hay de malo en uno de los dos potentes argumentos que entran en conflicto. El primer argumento es éste: si ustedes toman lo que está en las dos cajas, el ser casi con toda certeza lo habrá predicho y no habrá puesto el millón de dólares en la segunda caja; de manera que casi con toda seguridad ustedes sólo recibirán 1.000 dólares; mientras que si ustedes toman sólo lo que hay en la segunda caja, el ser casi con toda certeza lo habrá predicho y pondrá el millón de dólares en la segunda caja, de manera que casi con toda seguridad ustedes recibirán el millón de dólares. Por lo tanto, ustedes deberían tomar sólo lo que hay en la segunda caja. El segundo argumento es éste: el ser ya ha realizado su predicción y ha puesto ya, o no, el millón de dólares en la segunda caja. El millón de dólares ya está, o no, en la segunda caja, y lo que sea está ya fijado y determinado. Si el ser ha puesto ya el millón de dólares en la segunda caja, entonces si ustedes toman lo que hay en las dos cajas recibirán el millón de dólares más 1.000 dólares; mientras que si ustedes toman sólo lo que hay en la segunda caja, recibirán sólo el millón de dólares. Si el ser no ha puesto el millón de dólares en la segunda caja, entonces si ustedes toman lo que hay en las dos cajas, recibirán 1.000 dólares; mientras que si ustedes toman sólo lo que hay en la segunda caja, no recibirán nada. En cualquier caso, tanto si el millón de dólares ha sido depositado como si no, ustedes recibirán más dinero, 1.000 dólares más, tomando lo que hay en las dos cajas. (La acción de tomar lo que hay en las dos cajas, como se dice, *domina* a la acción de tomar sólo lo que hay en la segunda.) Por lo tanto, ustedes deberían tomar lo que hay en las dos cajas.

Desde que en 1969 presenté y discutí por vez primera este problema, se ha producido una abundante investigación y una iluminadora teorización sobre él.[1] En mi ensayo inicial, distinguía entre las probabilidades condicionales que señalan o acotan la *influencia* que una acción tiene sobre el estado que acontece y las meras probabili-

dades condicionales que no acotan tal influencia. Yo sugerí allí que, cuando entra en conflicto con el principio de dominación, no debería invocarse el principio de maximización de la utilidad condicional esperada si sus probabilidades condicionales fueran del segundo tipo (del de las que no acotan influencias). Argumentaba apoyándome en ejemplos intuitivos. (Esos ejemplos, en la medida en que intentaban incorporar cierta reflexividad, resultaban algo más complicados que los ejemplos discutidos por otros posteriormente.) Las predisposiciones genéticas, vinculadas entre sí, a contraer una enfermedad y a elegir una determinada carrera no deberían, argumentaba yo, llevar a alguien a evitar la elección de una carrera creyendo que eso aumentaría la estimación de su probabilidad de contraer la enfermedad: pues el que tenga esa estructura genética o el que acabe contrayendo la enfermedad no está *influenciado* o *afectado* por su elección de carrera. No se me ocurrió usar este asunto para el pleno y sistemático desarrollo de dos versiones competitivas de la teoría de la decisión, la causal y la evidencial, con sus distintas versiones del principio de utilidad esperada e incluso sus distintas versiones del principio de dominación.[2]

El principio tradicional de maximizar la utilidad esperada trata la utilidad esperada de una acción *A*, *UE(A)*, como la suma ponderada de las utilidades de sus posibles resultados (excluyentes) multiplicada por sus probabilidades, que suman 1.

$$UE(A) = \text{prob}(O_1) \times u(O_1) + \text{prob}(O_2) \times u(O_2) + \ldots + \text{prob}(On) \times u(On),$$
$$= \sum_{(i=1)}^{n} \text{prob}(Oi) \times u(Oi).$$

Un principio más adecuado, teniendo en cuenta que los resultados no tienen por qué ser probabilísticamente independientes de las acciones, multiplica la utilidad esperada no por las simples probabilidades de los resultados, sino por las probabilidades condicionales de los resultados dadas las acciones. Llamemos a eso la utilidad evidencialmente esperada de *A*, *UEE(A)*.[3]

$$UEE(A) = \text{prob}(O_1/A) \times u(O_1) + \text{prob}(O_2/A) \times u(O_2) + \ldots + \text{prob}(On/A) \times u(On),$$
$$= \sum_{(i=1)}^{n} \text{prob}(Oi/A) \times u(Oi).$$

Los teóricos causales de la decisión se sirven también no sólo de la probabilidad incondicional de un resultado, sino de una probabilidad que relaciona el resultado con la acción, esta vez no simplemente la probabilidad condicional, prob(O_1/A), sino alguna relación causal-probabilística que señala una influencia causal directa. La fórmula correspondiente a esas probabilidades causales incorpora la utilidad causalmente esperada de una acción *A*, *UCE(A)*.

A pesar de esta y otras elaboraciones técnicas —subjuntivos retrorrastreadores, explícita incorporación de cosquillas y metacosquillas, ratificabilidad de las decisiones, etc.—, y a pesar de los intentos de mostrar que el problema está irremediablemente mal definido o es incoherente,[4] la controversia sigue viva. Ninguna solución ha resultado completamente convincente.

El problema de Newcomb es un problema complicado, otros casos entrañan dificultades aún más complejas, el razonamiento en favor *de cada uno* de los argumentos en liza parece bastante concluyente —y nosotros somos criaturas falibles—. Sería irrazonable confiar absolutamente en alguna línea argumentativa particular cuando tales casos se presentan, o en algún principio particular de decisión.*

La cantidad en la primera caja, los 1.000 dólares, ha recibido poca atención.[5] Si el argumento de la dominación —el segundo argumento mencionado antes— es correcto, entonces ustedes harían mejor tomando lo que hay en las dos cajas aun si la cantidad en la segunda caja fuera mucho menor, un dólar por ejemplo, o incluso un céntimo, o una probabilidad de 1/10.000 de un céntimo. Sin embargo, pocos de nosotros escogerían ambas cajas en tal caso, no confiriendo *ningún* peso al otro argumento, según el cual si tomamos sólo lo que hay en la segunda caja, ganaremos casi con toda certeza el millón de dólares. Por otro lado, si el primer argumento mencionado más arriba es correcto, y se entiende como un argumento de utilidad esperada (sin necesidad de que las probabilidades condicio-

* Hace algunos años, en un seminario para graduados, varios estudiantes, y señaladamente David Cope, disputaban que alguien pudiera optar con certeza por la versión causal o por la versión evidencial de la teoría de la decisión, dados los poderosos argumentos de que dispone cada parte. Estoy en deuda con esas discusiones, pues me pusieron sobre la pista de mi actual línea de pensamiento. (Pero Howard Sobel me escribe que las cosas no son simétricas, pues sólo los teóricos causales han tratado de construir argumentos de su propia cosecha *y* de diagnosticar los [supuestos] errores de los argumentos encontrados, de acuerdo con el *desideratum* que yo propuse en su día en mi artículo.)

nales incrustadas expresen influencia de ningún tipo), entonces la cantidad de dinero X en la primera caja podría ser mucho mayor que los 1.000 dólares, y sin embargo, la persona no dejaría de elegir solamente lo que estuviera en la segunda caja. Supongamos que la probabilidad de que el ser prediga correctamente la acción de ustedes (para cualquier elección que ustedes puedan hacer) es 0,99. Siendo u la función de utilidad, la utilidad esperada de tomar sólo lo que está en la segunda caja es $0{,}99u$(del millón de dólares), mientras que la utilidad esperada de tomar lo que hay en las dos cajas es $u(X) + 0{,}01u$(del millón de dólares). En tal caso, la utilidad esperada de tomar sólo lo que está en la segunda caja será mayor que la utilidad esperada de tomar lo que hay en las dos cajas si $0{,}99u$(del millón de dólares) es mayor que $u(X) + 0{,}01u$(del millón de dólares) —esto es, si $0{,}98u$(del millón de dólares) es mayor que $u(X)$—. Suponiendo, entonces, que la utilidad crece linealmente con el incremento del dinero, la persona escogerá tomar sólo lo que hay en la segunda caja siempre que la cantidad de la primera caja sea inferior a 980.000 dólares. Así, por ejemplo, en un problema de elección que tuviera una estructura idéntica a la del problema de Newcomb, pero en el que la primera caja contuviera 979.000 dólares y la segunda caja, lo mismo que antes, un millón de dólares o nada, la persona no tomaría el contenido de ambas cajas, sino sólo de la segunda caja. Claro que la utilidad del dinero no crece linealmente con su incremento dentro de esta gama, pero eso no nos desvía aquí demasiado de nuestros propósitos (es la utilidad de $M + X$ la que crecerá proporcionalmente menos que su cantidad de dinero). La noción general, sin embargo, se mantiene; para grandes cantidades de dinero en la primera caja, 900.000 dólares por ejemplo, suponiendo que el ser es muy preciso en sus predicciones, quien propusiera el primer argumento tomaría sólo lo que hay en la segunda caja. No obstante, pocos de nosotros se sentirían confortables secundando el primer argumento en este caso y quitándole *toda* fuerza al otro argumento, según el cual haremos mejor en ambos casos tomando lo que hay en las dos cajas.

Variando la cantidad de dinero en la primera caja, podemos hacer que la gente se sienta muy incómoda con su argumento predilecto para elegir en el problema inicial de Newcomb. La gente que inicialmente elige las dos cajas no está dispuesta a secundar el argumento de la dominación cuando la cantidad en la primera caja desciende hasta un dólar; la gente que inicialmente elige sólo la segunda caja no está dispuesta a secundar el argumento de la utili-

dad esperada (con probabilidades condicionales que no acotan influencias) cuando la cantidad en la primera caja se eleva a 900.000 dólares. Eso sugiere que nadie tiene una confianza *total* en el argumento que secunda inicialmente en el problema de Newcomb. Nadie está dispuesto a aplicar sin reservas y sin fronteras el argumento que parecía moverle en este caso.

Podría ocurrir que una persona pusiera distintas dosis de confianza en varios principios de decisión (y en los argumentos que van con ellos). Podemos por el momento reducirnos a esos dos principios de maximización de la utilidad (condicionalmente) esperada formulados, respectivamente, por la teoría causal y por la teoría evidencial de la decisión. Podríamos representar esas diferentes dosis de confianza con grados de confianza entre 0 y 1 (inclusive) que sumen 1, o con grados que no sumen 1 (dejando abierta la posibilidad de que ambos principios sean incorrectos en un caso dado), o con ponderaciones de confianza que no fueran grados entre 0 y 1. Para una determinada persona, sea *Wc* el peso que ella confiere al principio de utilidad esperada de la teoría causal de la decisión, y sea *We* el peso que confiere al principio de la utilidad esperada de la teoría evidencial de la decisión. Sea *UCE(A)* la utilidad causalmente esperada de una acción *A*, la utilidad de esa acción, tal como sería calculada de acuerdo con (alguna versión favorecida de) la teoría causal de la decisión; sea *UEE(A)* la utilidad evidencialmente esperada de la acción *A*, la utilidad de esa acción tal como sería calculada de acuerdo con la teoría evidencial de la decisión. Vinculado a cada acción habría un valor decisional, *VD*, un valor ponderado de su utilidad causalmente esperada y de su utilidad evidencialmente esperada, y ponderado de acuerdo con la confianza de esa persona en la orientación que le proporciona cada uno de esos dos tipos de utilidad esperada.

$$VD(A) = Wc \times UCE(A) + We \times UEE(A)$$

Y la persona tiene que elegir una acción con máximo valor decisional.[6]

Sugiero que se pueda ir más lejos, y decir no sólo que no tenemos certeza respecto de *cuál* de esos dos principios, *UCE* y *UEE* es (tomado en exclusiva) correcto, sino que ambos principios son legítimos y debe darse a cada uno el peso que le es debido. Los pesos, pues, no son medidas de incertidumbre, sino medidas de la fuerza legítima de cada principio. Tenemos entonces una teoría *normativa*

que guía a las personas a elegir una acción que maximice el valor decisional.

Si un maximizador de valor decisional confiere pesos distintos de cero a *Wc* y a *We*, se verá inducido a modificar su elección en el problema de Newcomb: de una caja a las dos, cuando la cantidad en la primera caja aumenta suficientemente; de dos cajas a una, cuando la cantidad en la primera caja desciende suficientemente. Cuando se trata de maximizadores de valor decisional, esos cambios resultan perfectamente predecibles. (De manera que la maximización *VD* tiene consecuencias conductuales cualitativas y contrastables, al menos en el caso de quienes se atienen a esta teoría normativa.)

Hay varias estructuras matemáticas que podrían conferir un papel a *UCE* y a *UEE*, pero la fórmula *VD* es especialmente simple, y sería prematuro atender ahora a cosas más complicadas. Obvio es decir que la estructura *VD* ponderada no tendrá, por sí misma, mucho valor orientativo para nadie. ¿Cuáles deberían ser exactamente los pesos conferidos? ¿Debe una persona usar los mismos pesos en todas las situaciones decisionales, o podrían acaso variar los pesos según los distintos tipos de situaciones decisionales? ¿O variar, más sistemáticamente, según donde caiga una situación decisional a lo largo de alguna dimensión *D* —cuanto más a la izquierda, tanto más plausible el uso de uno de los criterios de decisión (y por consiguiente, tanto mayor el peso que se le confiere), cuanto más a la derecha, tanto más plausible el uso del otro—? Tan enhorabuena vendría una teoría que determinara o restringiera los pesos, como una teoría que determinara o restringiera las probabilidades previas dentro de una estructura bayesiana, o una que determinara o restringiera el contenido substantivo de las preferencias en el marco de los axiomas usuales de orden. Aun así, en cada uno de estos casos, la estructura general puede resultar iluminadora.

Que hay que conceder algún peso a ambos factores, *UCE* y *UEE*, significa que *UEE* recibirá algún peso aun en decisiones acerca de casos en los que no se constata influencia causal alguna de la acción sobre el resultado relevante, por ejemplo, en casos en los que la elección de una carrera señala (pero no afecta a las) diferentes probabilidades de contraer, o de haber ya contraído, una enfermedad terrible. En mi viejo artículo consideré absurdo conferir el menor peso a esas consideraciones. Y sin embargo, no ignoraba que el componente evidencial de la fórmula *VD* había tenido las mayores consecuencias sociales en la historia humana, como lo atestigua la historiografía sobre el calvinismo y el papel que su concepción

de los *signos* (que no las causas) de elección desempeñó en el desarrollo del capitalismo. (Puede ser una consecuencia causal de una acción el que la persona crea —y le haga feliz tal creencia— algo que es señalado por esa acción, pero no es causado por ella. Mas alguien que introdujera esto como razón para emprender la acción debería cuidarse de apelar a esas consecuencias felices como razón para mantener la creencia.)[7]

Los teóricos de la racionalidad no han dejado de perseguir la formulación del conjunto correcto y completo de principios susceptibles de ser aplicados sin reservas a todas las situaciones decisionales. Pero aún no lo han conseguido —y en cualquier caso, no tenemos demasiada confianza en que lo consigan—. Dado lo cual, ¿no debería un individuo racional y prudente defender sus apuestas? Es más: ninguno de los principios resulta por sí solo plenamente adecuado. No se trata simplemente de que aún estemos aguardando el argumento clave en favor del principio correcto. No digo que el marco del valor decisional baste para poner de acuerdo a todos los teóricos de la decisión. Éstos seguirán discrepando acerca de los pesos que hay que asignar a los distintos principios de decisión, aun cuando coincidan en los principios que deberían incluirse. Es esta discrepancia respecto de los pesos lo que explica las distintas elecciones en el problema de Newcomb, pero el hecho de que confiramos *pesos* (en vez de atenernos meramente a un principio) explica la oscilación de las decisiones a medida que variamos la cantidad en la primera caja. La estructura *VD* representa el hecho de que cada principio *UEE* y *UCE* capta razones legítimas (de un tipo), y no estamos dispuestos a dejar completamente de lado ninguno de los tipos.[8]

Resulta un tanto extraño que los teóricos de la decisión hayan sabido mostrar tal confianza en sus concepciones. Pues, si formulamos el asunto mismo del principio correcto de decisión como un problema de decisión, como un problema de decisión acerca de qué principio de decisión debería ser secundado[9] (podríamos imaginar que se han inventado las píldoras que pueden transformarnos en seguidores consistentes de cada principio), entonces no resulta obvio qué dirán los principios de decisión en liza. En particular, no resulta obvio si cada uno de ellos se propondrá a sí mismo como la alternativa preferida. Eso dependerá de cómo sea el mundo. Si el mundo ofrece muchas situaciones parecidas a la del problema de Newcomb, con beneficios considerables, entonces se puede predecir que tomar la píldora *UEE* llevará a mejores consecuencias *causales*, de manera que el principio *UCE* recomendará tomar la píldora *UEE*, no la píldora *UCE*. Si, en cambio, el mundo fuera a suministrar muchas

situaciones parecidas a la de mi ejemplo de la enfermedad u otras muchas similares (el ejemplo de Salomón, construido por Gibbard y Harper, etc.),[10] entonces la persona que secundara el principio *UEE* (sin ninguna adición de «cosquillas»), dejaría escapar con frecuencia importantes ganancias (a causa de las desgracias que anuncian). Puesto que eso puede predecirse, el principio *UEE* mismo recomendaría la ingesta de la píldora *UCE* como una acción con mayor utilidad *UEE* que la ingesta de la píldora *UEE*. (En este caso, el principio *UCE* recomienda también la ingesta de la píldora *UCE*. ¿Hay algún ejemplo en el que la *UEE* de ingerir la píldora *UCE* sea mayor —de manera que el principio *UCE* lo recomienda—, aun cuando la *UCE* de ingerir la píldora no sea mayor, y por lo tanto, el principio *UCE* no recomiende su ingesta? En este caso, las dificultades vendrían en tropel.) Así como no hay ninguna política inductiva determinada, ninguna función-*c* carnapiana, que sea la mejor o la más efectiva independientemente de la urdimbre del mundo, quizá tampoco haya ningún principio único óptimo de decisión racional.[11] Y así como queremos dar a nuestros procedimientos inductivos un margen para el aprendizaje, para incorporar parámetros que habrán de ir definiéndose al ganarse en experiencia mundana, así también queremos que nuestros principios de decisión racional incorporen parámetros que puedan ir definiéndose para adaptarse a la urdimbre que va descubriéndose en el mundo en el que se toman las decisiones. (En ambos casos, la evolución podría haber cargado ya con una parte significativa de la tarea de fijar los parámetros que se adaptan al mundo real, pero esto no significa que debamos esperar que nuestras particulares políticas inductivas, o nuestros particulares principios de decisión, sean aplicables en cualquier situación imaginable de ciencia ficción, ni menos que debamos considerarlos como válidos *a priori*.) El marco de los valores de decisión, incorporando pesos susceptibles de modificación a lo largo del tiempo, es una de las vías por las que podría lograrse una adaptación al mundo real.

El valor decisional que hemos definido se basaba en los componentes *UCE* y *UEE*, pero cualquier principio alternativo plausible de decisión, o cualquier factor en la decisión, podría añadirse como un término que trajera consigo su propio peso. En particular, podríamos añadir a la fórmula la utilidad simbólica de una acción, su *US* (que incorpora la utilidad de los varios resultados y acciones *simbolizados* por la acción), la cual traería consigo su propio peso, W_S. (Lo mejor es no tratar de incorporar la utilidad simbólica a otras utilidades, porque podría ser que no se rigiera por una fórmula de

valor esperado y porque podríamos querer mantener una pista separada para la utilidad simbólica, pensando que resulta apropiado conferir a este factor diferentes pesos en diferentes tipos de situaciones de elección.) La fórmula para el valor decisional de *A*, *VD(A)*, resultaría entonces:

$$VD(A) = Wc \times UCE(A) + We \times UEE(A) + Ws \times US(A)$$

Sería instructivo investigar las características formales de esta estructura del valor decisional. No resultaría sorprendente que este principio de combinación ponderada, lo mismo que otros criterios ya investigados por la literatura sobre decisión en condiciones de incertidumbre, no consiguiera a veces exhibir determinados rasgos deseables.[12]

La utilidad simbólica no es un tipo distinto de utilidad, un tipo que mantuviera con la utilidad estándar una relación parecida a la que mantiene el significado metafórico con el literal. Ocurre, más bien, que la utilidad simbólica representa un tipo distinto de *conexión* —simbólica— con el tipo normal de utilidad. Un tipo de conexión que se da junto a las consabidas conexiones causal y evidencial. La utilidad simbólica de una acción *A* está determinada por el hecho de que *A* tiene conexiones simbólicas con resultados (y acaso con otras acciones), los cuales poseen ellos mismos el tipo estándar de utilidad, lo mismo que la *UCE* de *A* está determinada por las conexiones causal-probabilistas de *A* con resultados que poseen la utilidad estándar.*

¿Deberíamos asegurarnos de que estos tipos de conexión —causal, evidencial y simbólica— son excluyentes? La primera fórmula para *VD* la definía como una suma ponderada de *UCE* y *UEE*. Sin embargo, la *UEE* de una acción incluye sus componentes causales, puesto que las probabilidades condicionales de los resultados dadas las acciones, prob(*O/A*) que el teórico de la *UEE* utiliza, incorpora influencias causales, si existen. En nuestra fórmula de suma

* Por lo tanto, es necesario imponer una condición adicional a la situación estándar, discutida en la nota al pie de las páginas 60-61, para medir la utilidad. Esa situación debe ser de tal naturaleza, que las acciones no tengan conexiones simbólicas o evidenciales relevantes con resultados de utilidad. La utilidad ha de ser calibrada en contextos causales en los que se secunda un principio de valor esperado, y las utilidades (saneadas) así descubiertas han de utilizarse en situaciones en las que las acciones tienen también conexiones evidenciales y simbólicas con los resultados valorados. Sin embargo, el valor de estos últimos resultados se mide en situaciones que son plenamente causales.

ponderada, pues, ¿no deberíamos interpretar la *UEE* como la utilidad esperada representada por aquellas (porciones de) probabilidades que *no* son (simplemente derivativas de) las probabilidades causales? Y análogamente, ¿no debería la utilidad simbólica *US* de una acción ser su utilidad simbólica que no es (simplemente) derivativa de, y está representada en, aquellas conexiones causales y evidenciales?[13]

¿Deberíamos incluso incorporar otros componentes adicionales a la estructura *VD*? Se podría sugerir la inclusión explícita de un componente que tuviera que ver con el modo en que una acción casa con la imagen que de sí propia tiene la persona, con el modo en que esa acción expresa el yo de esa persona. De hecho, nuestros tres componentes cubren ya buena parte de ese territorio. Aun cuando realizar una acción del tipo que sería realizado por una determinada clase de persona acaso no *cause* que el agente se convierta en esa clase de persona, quizá simbolice su ser de esa manera, quizá constituya algún tipo de evidencia de que lo es y quizá tenga la consecuencia causal de hacerle más fácil el mantenimiento de una *imagen* de sí mismo como alguien que es de esa clase de persona. Esto último constituye una consecuencia causal real de una acción, y puede revestir una utilidad considerable. De aquí que este tipo de consecuencia —el modo en que la realización de una determinada acción afecta a la imagen que la persona tiene de sí misma— pueda desempeñar un papel destacado, explícito, en una teoría explicativa de la conducta de esa persona, aun cuando se trate de un tipo de consecuencia que el agente mismo no puede tomar fácilmente en cuenta de una manera explícita («voy a hacer *A* para facilitarme el mantenimiento de la imagen de mí mismo como persona del tipo *K*») sin menguar con ello el efecto mismo.[14] No deberíamos interpretar la expresividad como si quedara agotada por esas otras categorías independientes, estrechamente concebidas. Pues, como ya ha dicho, las categorías de lo simbólico y de lo expresivo van entrelazadas.

La alineación de la categoría de lo linguístico junto a las categorías de que ya disponemos, del modo siguiente: causal/evidencial/lingüístico/simbólico, sugiere dos cuestiones. ¿Cómo difiere la naturaleza de lo simbólico de la de lo meramente linguístico (lo simbólico se pega más a la utilidad retroactivamente imputada, pero no necesariamente en cada momento)? ¿Y cómo puede surgir lo linguístico a partir de lo causal y evidencial? Cuando las conexiones causales y evidenciales (que surgen de una estructura ramificada de regularidades causales y estadísticas) son de conocimiento común, alguien podría producir intencionalmente un signo evidencial de *p* para ha-

cer creer a otro ese *p*. Esto representaría un paso crucial más allá del significado natural de Grice, que es un signo evidencial, e implicaría un desarrollo intencional capaz de generar en otro una creencia, es decir, recorrer una parte del camino hacia un significado no-natural (en el sentido de Grice), en el que se pretende que esa intención se reconozca como tal.[15] Podría producirse ese signo evidencial con objeto de inducir a una creencia en un *p* verdadero que la otra persona no puede observar correctamente de un modo independiente en estos momentos. Pero también —y acaso no menos probablemente— podría haberse producido primero ese signo con objeto de engañar a otra persona para que creyera *p*, sembrando evidencias de que *p* es verdadero, cuando, en cambio, es falso. El primer enunciado que valió por alguna otra cosa bien pudo haber sido una mentira, un signo natural fingido. Si es verdad que el lenguaje define a la humanidad, expresando las capacidades racionales humanas y distinguiéndolas de las de los animales, podríamos estar asistiendo a una curiosa vuelta de tuerca de la doctrina según la cual hemos nacido en el pecado original.

EL DILEMA DEL PRISIONERO

El dilema del prisionero es una situación harto discutida, en la que la selección a que cada parte procede de una acción (robustamente) dominante, que aparece como la única racional, deja a cada una de ellas peor que si hubieran elegido la acción dominada, más cooperativa. La combinación de (lo que parecen ser) sus racionalidades individuales les lleva a dejar pasar una situación accesible mejor, y es, por lo tanto, paretianamente subóptima.

La situación general debe su nombre a un caso particular de la misma: un inspector ofrece a dos personas encarceladas en espera de juicio las siguientes opciones. (La situación entre los prisioneros es simétrica; no pueden comunicarse para coordinar sus acciones en respuesta a la oferta del inspector, o si pueden, no tienen manera de hacer prevalecer el acuerdo al que podrían llegar.) Si un prisionero confiesa y el otro no, el primero no va a la cárcel, y el segundo recibirá una sentencia de doce años de cárcel; si ambos confiesan, cada uno de ellos recibirá una sentencia de diez años; si ninguno de los dos confiesa, cada uno de ellos recibirá una sentencia de dos años. El cuadro 2 representa la situación a la que se enfrentan, representando las entradas en la matriz el número de años que aguardan en prisión al primer y al segundo prisionero, respectivamente.

		Prisionero II	
		No confesar	Confesar
Prisionero I	No confesar	2,2	12,0
	Confesar	0,12	10,10

CUADRO 2

Cada prisionero razona como sigue: «Si la otra persona confiesa y yo no, a mí me caerán doce años de cárcel; mientras que si yo confieso, me caerán diez años. Si la otra persona no confiesa y yo no confieso, me caerán dos años de cárcel; mientras que si yo confieso, me dejarán libre. En cualquier caso, haga lo que haga la otra persona, a mí me irá mejor confesando que no confesando. Por lo tanto, confesaré». Cada prisionero razona del mismo modo: ambos confiesan y a ambos les caen diez años de cárcel. Si ninguno de los dos hubiera confesado, les habrían caído sólo dos años de cárcel. Las racionalidades individuales combinan entre sí para producir un mixto conjunto. Y la situación es estable en el siguiente sentido: ninguna de las partes tiene incentivos para realizar la otra acción (más cooperativa), dado que la otra parte confesará. Las acciones de confesar están en equilibrio.

La situación de dilema del prisionero es un caso de una estructura más general (véase el cuadro 3), en la que cada parte puede elegir entre dos acciones —llamémosles *D*, la acción dominante, y *C*, la cooperativa— y tiene las siguientes preferencias respecto de los posibles resultados *a, b, c* y *d* de las acciones combinadas. La persona I prefiere *c* a *a, a* a *d* y *d* a *b,* mientras que la persona II prefiere *b* a *a, a* a *d* y *d* a *c*. Puesto que la persona I prefiere *c* a *a* y *d* a *b,* la acción *D* domina a la acción *C,* y elige *D*. Puesto que la persona II prefiere *b* a *a,* y *d* a *c,* la acción *D'* domina a la acción *C',* y elige *D'.* Juntando *D* y *D'* sale el resultado *d,* aunque ambas prefieren el resultado *a* (que resultaría de *C* y *C'*) al resultado *d*. Por tanto, estos simples hechos acerca de la estructura de la matriz 2 × 2 y de la estructura del orden de preferencias de cada persona parecen bastar para configurar una situación de dilema del prisionero.

		II	
		C'	D'
I	C	a	b
	D	c	d

CUADRO 3

Algunos han sostenido que una persona racional en esta situación, sabiendo que la otra persona es también una persona racional que sabe tanto de la situación como ella misma, se percatará de que cualquier razonamiento que le resulte a sí misma convincente le resultará asimismo convincente a la otra. De modo que si concluye que la acción dominante es la mejor, la otra persona lo concluirá también; si concluye que la acción cooperativa es la mejor, también la otra lo hará. En esta situación, lo mejor sería entonces concluir que la acción óptima es la cooperativa, y, percatada de todo eso, la otra persona concluirá (de un modo u otro) en el mismo sentido. Este tipo de argumentación ha tenido una recepción mixta.

El dilema del prisionero tiene paralelismos con el problema de Newcomb, sean o no ambos idénticos en todos sus rasgos esenciales (como algunos han sostenido). Ambos entrañan dos argumentos que llevan a diferentes acciones: un argumento se funda en el principio de dominación, interpretado de un modo congenial con la teoría causal de la decisión; el otro argumento se funda en la consideración de lo que cada acción indicaría (y por lo tanto, en el resultado por el que habría que apostar) de un modo congenial con la teoría evidencial de la decisión. El argumento, según el cual, en el dilema del prisionero, ustedes deberían esperar que la otra persona hiciera lo mismo que ustedes, aun cuando la acción de ustedes no afectara causalmente a las de la otra persona, se compadece bien con el principio de la maximización de la utilidad evidencialmente esperada, principio en el que las probabilidades condicionales no representan necesariamente alguna influencia causal. La teoría causal de la decisión recomienda ejecutar la acción dominante; la teoría evidencial de la decisión, recomienda ejecutar la acción cooperativa si ustedes creen que la otra parte es relevantemente parecida a ustedes. No es necesario que ustedes tengan la certeza de que los dos actuarán de modo parecido; basta con que las probabilidades condicionales de las acciones de la otra parte, dadas las de ustedes, varíen suficientemente. (Obsérvese también que la teoría evidencial de la decisión podría llegar a recomendarles la acción dominante si ustedes creen que es probable que la otra parte ejecute una acción *distinta* de la suya, o simplemente si la acción de la otra parte es independiente de la de ustedes, pero ustedes imputan probabilidades suficientemente pesimistas a la inclinación de la otra parte a cooperar.) Como ocurría en el problema de Newcomb, nuestra confianza en estas distintas posiciones seguramente no es completa, y quizá deseemos darle a cada una de ellas algún peso legitimado.

En el caso del problema de Newcomb, la plurigarantía de peso

legitimado (o, alternativamente, la falta de confianza completa) se revelaba en la oscilación de las decisiones cuando se procedía a variar la cantidad de dinero en la primera caja. (Sin embargo, la estructura del problema permanecía constante juzgada de acuerdo con los dos principios de decisión en liza, cada uno de los cuales habría mantenido la misma decisión a través de esos cambios.) En el caso del dilema del prisionero, la cuestión es qué deberían hacer agentes racionales imbuidos del conocimiento común de que se enfrentan a agentes racionales. Los propugnadores de los dos argumentos en disputa piensan que la estructura abstracta del cuadro 3 basta para inclinar concluyentemente la balanza del lado de su argumento predilecto. Todo lo que necesita el argumento de la dominación es que la persona I prefiera *c* a *a* y *d* a *b*, y que la persona II prefiera *b* a *a* y *d* a *c*. Todo lo que parece necesitar el argumento de la utilidad evidencialmente esperada acerca de los agentes racionales es que cada uno de ellos tenga conocimiento común de que cada uno es un agente racional y que todos y cada uno prefieren *a* a *d*. Sin embargo, a falta de una completa confianza en esos argumentos por parte de la gente, deberíamos encontrarnos con que las variaciones en las cantidades* de los beneficios dentro de la estructura abstracta del cuadro 3 (aun manteniéndose el *orden* de las preferencias de las personas) producirá cambios en las decisiones que tomaría la gente.

Supongamos que la utilidad se mide en una escala de intervalo, única a lo largo de una transformación lineal positiva, con una unidad y un punto cero arbitrarios, de acuerdo con alguna variante de los axiomas estándar de von Neumann-Morgenstern.[16] En la situación representada por el cuadro 4,

		II	
		C'	D'
I	C	1,000, 1,000	0, 1,001
	D	1,001, 0	1,1

CUADRO 4

en donde las entradas de la matriz son números de utilidad tales, que pensaríamos que la cooperación es una elección racional. En

* Más exactamente —puesto que la utilidad se mide en una escala de intervalo—, en las *ratios* de las diferencias de cantidades. Aunque la discusión que sigue prescinde con frecuencia de esta complicación en interés de la claridad, siempre podría ser convenientemente reformulada para incluirla.

general, cuando los beneficios de la solución cooperativa son mucho más altos que los de la solución dominante, y cuando los beneficios asociados a las acciones que no intersectan ofrecen sólo ganancias menores, o aun pérdidas, respecto de estas dos, entonces pensaremos sin vacilar que la cooperación es racional y hallaremos que el argumento de la dominación tiene poca fuerza. En cambio, en el cuadro 5,

		II	
		C'	D'
I	C	3,3	—200, 500
	D	500, —200	2,2

CUADRO 5

la solución cooperativa es sólo ligeramente mejor que la dominante, y los valores extremos en las ganancias para las acciones que no intersectan divergen drásticamente. Si no guardamos especiales vínculos con la otra parte, o si carecemos de conocimiento concreto sobre las probabilidades de la acción de la otra parte, entonces pensaremos que en la situación del cuadro 5 lo racional es realizar la acción dominante, no correr el riesgo de que la otra parte realice su acción dominante, para lo cual se ve fuertemente incentivada. (Y si prosigo el razonamiento y pienso que la otra parte es probablemente como yo, entonces puedo elegir tranquilamente la acción dominante, confortado por el pensamiento de que la otra parte hará lo mismo.)

Esas oscilaciones en la decisión que uno tomaría, oscilaciones que dependen de las (*ratios* de las diferencias en las) particulares entradas de utilidad numérica en la matriz, concuerdan con el anterior principio de maximización del valor decisional por parte de quienes confieren algún peso a cada uno de los particulares principios *UCE* y *UEE*. En qué punto preciso oscilará la decisión de una persona a medida que varían las utilidades, dependerá de su confianza en cada uno de estos principios (esto es, de qué pesos les asigna implícitamente), así como también de las probabilidades que asigna a la posibilidad de que la acción de la otra persona sea la misma que la suya. Obsérvese, no obstante, que aun si a esto último se le asigna una probabilidad 1, y aun si el agente confiere mayor peso al principio *UEE* que al *UCE*, no necesariamente realizará la acción cooperativa. Si las cotas de utilidad son suficientemente altas y casan con la situación del cuadro 5, ese hecho puede combinarse

con el peso conferido al principio *UCE*, o con el mismo principio de dominación (en su variante causal), o con algún otro principio que confiera mayor peso al nivel de seguridad, para que resulte recomendada la acción dominante. Ni siquiera la absoluta confianza en que la otra persona actuará como ustedes basta para garantizar que ustedes realizarán la acción cooperativa —a falta de una confianza absoluta, o de un peso absoluto conferido al principio *UEE*—.* (Hasta ahora he partido del supuesto de que lo que aplica una persona en todas las situaciones decisionales es una particular versión del principio *VD*, con los particulares pesos por él conferidos. Podría darse el caso, sin embargo, de que, para un conjunto dado de principios constituyentes de decisiones, una persona asignara a esos principios pesos diferentes según la situación decisional a la que se enfrentara. Con todo, cualquier tipo de situación en la que más de un principio particular recibiera un peso positivo sería susceptible de encajar en alguna que otra estructura *VD*.)

En la sección anterior, incorporamos la utilidad simbólica de realizar una acción, su *US*, a la estructura *VD*, junto a la *UCE* y a la

* «Mas ¿no *insistiría* una teoría correcta en que cuando la probabilidad de que la otra persona actúe de la misma forma que ustedes es igual a 1, ustedes *deberían* elegir la acción cooperativa en la situación de dilema del prisionero, cualquiera que sea la magnitud de las diferencias de utilidad en la matriz?» Podríamos, sin embargo, preguntarnos si la persona (un nivel más arriba) tiene completa confianza en su estimación de la probabilidad 1, y si a falta de una confianza completa se verá afectada su acción en esta situación altamente arriesgada. Véase Daniel Ellsberg, «Risk, Ambiguity, and the Savage Axioms», *Quarterly Journal of Economics* 75 (1961): 643-669.

Obsérvese también que el argumento pasa demasiado apresuradamente de (1) la común racionalidad, a (2) las partes harán lo mismo, a (3) tachar la casilla superior derecha y la casilla inferior izquierda de la matriz, casillas que representan acciones distintas, a (4) argumentar que, dado que la elección tiene que hacerse entre las otras dos casillas restantes, ambas partes deberían elegir la casilla preferida por cada una de ellas, es decir, ambas deberían realizar la acción cooperativa. El supuesto del común conocimiento de la racionalidad nos permite suponer que razonaremos de la misma forma y acabaremos haciendo lo mismo. Pero quizá *esto* resulte del razonamiento que cada persona haga atendiendo a las cuatro casillas de la matriz, concluyendo cada una de ellas que, a la vista de la situación estratégica global presentada por la matriz completa, con sus cuatro casillas, yo (o él, o ella) debería abstenerme de la acción cooperativa, de manera que las dos personas acabarían en la casilla inferior derecha, no cooperativa —satisfaciendo *así* la condición de actuar idénticamente—. Saber por anticipado que acabaremos haciendo lo mismo significa que sabemos que no acabaremos en la casilla superior derecha ni en la casilla inferior izquierda. Pero eso no significa que podamos empezar eliminándolas, para luego razonar sobre la situación restante. Pues acaso el razonamiento que nos lleva a *acabar* ejecutando la misma acción depende de que no *empecemos* por eliminar esas esquinas divergentes.

UEE. Podría pensarse que el que una acción *tenga* utilidad simbólica se revela por sí mismo *completamente* en las entradas de utilidad de la matriz para esa acción (por ejemplo, quizá cada una de las entradas es alzaprimada por una determinada cantidad fija que vale por la utilidad simbólica de la acción), de manera que no habría necesidad de ningún factor de *US* separado. Sin embargo, el valor simbólico de una acción no está determinado solamente por *esa* acción. El significado de una acción puede depender de que haya otras acciones disponibles, de sus beneficios y de las acciones a disposición de la otra parte o de las otras partes. Lo que la acción simboliza, lo simboliza en cuanto que realizada en *esta* particular situación, y preferida a *estas* alternativas particulares. Si una acción simboliza «ser una persona cooperativa», tendrá ese significado no simplemente porque acarrea los dos posibles beneficios que acarrea, sino también porque ocupa una determinada posición dentro de la matriz bipersonal —es decir, ser una acción dominada que (unida a la acción dominada de otra persona) arroja para cada uno un beneficio más elevado que la combinación de las acciones dominantes. De aquí que su *US* no sea una función de los rasgos que pueden captarse considerando esa acción aislada, proyectando simplemente los estados en las consecuencias.[17] El valor simbólico de una acción depende de la entera decisión, de toda la matriz del juego. No queda adecuadamente representada por medio de alguna adición o de alguna substracción de las utilidades de las consecuencias *dentro* de la matriz. Algunos autores suponen que *cualquier cosa* puede ser formalmente construida y encajada en las consecuencias,[18] por ejemplo, cómo *siente* uno el haber realizado una acción, el hecho de haberla ejecutado, o el hecho de que caiga bajo determinados principios deontológicos. Pero si las *razones* para realizar una acción *A* afectan a su utilidad, entonces tratar de construir y encajar esa utilidad de *A* en sus consecuencias tiene que acabar alterando esa acción y cambiando las razones para hacerla; mas la utilidad de *esa* acción alterada dependerá de las razones para hacerla, lo que volverá a alterar la acción, y así sucesivamente. Por lo demás, las utilidades de un *resultado* pueden cambiar si la acción se hace por determinadas razones.[19] Lo que queremos que representen las utilidades de los resultados, por lo tanto, son las utilidades *condicionales* de los resultados, dado que la acción se ha ejecutado por determinadas razones.* Esto le crea un problema al consecuencialismo a la

* O incluso la utilidad condicional del resultado, dado que la acción se ha ejecutado por ciertas razones *y lleva al resultado*. En la literatura económica sobre las subastas, se señala que la estimación del valor del resultado por parte de una per-

hora de tratar asuntos de consistencia dinámica; pues podría darse el caso de que el hecho de haber llegado a un determinado subárbol del árbol decisional les proporcionara a ustedes una información que alterara la utilidad de un resultado futuro. Si tratamos de lidiar con este problema insistiendo en que las utilidades dentro del árbol sean siempre utilidades condicionales plenamente definidas, entonces no podremos tener los *mismos* resultados en dos sitios distintos cualquiera del árbol decisional —lo que socava la formulación de principios normativos generales encargados de gobernar ese árbol—. (Para *cada* hecho acerca de una acción, podría haber una descripción que les permitiera a ustedes clasificar ese hecho como una consecuencia de la acción, pero de aquí no se sigue que haya una descripción tal, que esa descripción incorpore *todos* los hechos acerca de la acción a las consecuencias de la acción. El orden de los cuantificadores importa.)

Estas reflexiones muestran que, en las situaciones de dilema del prisionero, debería entenderse que una acción tiene utilidad por sí misma, no simplemente por entrañar una adición constante de utilidad *dentro* de una columna de su matriz.[20] Pero yo quiero sostener algo más fuerte, a saber: que esa utilidad es una utilidad *simbólica*. No me refiero simplemente al tipo corriente de utilidad aplicada a una acción en vez de a un resultado. La utilidad de que hablo implica un tipo distinto de conexión. En algunas situaciones de dilema del prisionero, la realización de la acción dominada —habitualmente conocida como «acción cooperativa»— puede tener un valor simbólico para la persona. Puede valer por ser ella misma una persona cooperativa en interacciones con otros, un participante voluntario y sin reparos en empresas conjuntas de mutuo beneficio. El

sona podría cambiar cuando ésta descubre que su particular oferta era la oferta ganadora, siempre que esto sea indicio de que otros licitantes de los que se tiene noticia habrían tenido información o llegado a conclusiones que les hicieron valorar el resultado menos que él. La literatura sobre ratificabilidad observa que el que «yo decida hacer A» puede afectar a la estimación de la probabilidad de una consecuencia C de A, en donde la prob(C/decido hacer A) no es igual a la prob(C); mientras que, en cambio, la literatura sobre subastas insiste en que «mi hacer A tiene éxito en conseguir C» puede afectar a la utilidad de C, quizá alterando las probabilidades de otra información que afecta a la utilidad de C. Así, una teoría de la decisión completamente formulada no sólo debe servirse de la utilidad condicional —véase mi *The Normative Theory of Individual Choice* (1963; reimp. Nueva York: Garland Press, 1990), págs. 144-158—, sino la teoría de la utilidad que se emplee no puede ser simplemente u(resultado O/se hace la acción A), sino u(resultado O/se hace la acción por las razones R, y esta A hecha por R lleva a O).

cooperar en tal situación puede entonces agruparse con otras actividades cooperativas que no están incrustadas en situaciones de dilema del prisionero. De aquí que no cooperar en esta particular situación de dilema del prisionero pudiera llegar a amenazar la cooperación de la persona en esas otras situaciones —la línea que las separa acaso no esté tan marcada, y la motivación de la persona para cooperar en las otras situaciones quizá sea también parcialmente simbólica—. Puesto que la persona atribuye una gran utilidad a ser una persona cooperativa, en una determinada situación del dilema del prisionero realiza la acción dominada que simboliza eso.*

Esto no significa que la persona se limite a atender a la *US* de esa acción. También tendrá en cuenta las particulares entradas de utilidad de la acción, y el modo en que son evaluadas por los principios *UCE* y *UEE.* El valor decisional que la acción revista para la persona dependerá de estas tres cosas —*US, UCE* y *UEE* de la acción— y del peso que les conceda. Así, el mero hecho de que confiera alguna utilidad simbólica (positiva) a ser una persona cooperativa no garantiza que acabe realizando la acción cooperativa en todas las situaciones de dilema del prisionero.

No digo que el único significado simbólico posible relevante para las situaciones de dilema del prisionero sea «ser una persona cooperativa». Alguien podría pensar que realizar la acción *dominante* en tales situaciones simboliza «ser racional, no dejarse arrastrar por el sentimentalismo». Pensando que esto es bastante importante, conferirá gran utilidad simbólica (dentro de su estructura *VD*) a realizar la acción dominante, sumándose esa utilidad al peso que él mismo concede a la *UCE* o al principio de dominación. Algunos investigadores del problema de Newcomb que propugnan la racionalidad de tomar lo que hay en las dos cajas superan la incomodidad que les produce el hecho de que a ellos y a quienes como ellos piensan les va peor en este problema que a los maximizadores de *UEE* diciéndose que su «moral» es: «Si alguien es muy bueno prediciendo la conducta y recompensa abundantemente la irracionalidad predicha, entonces la irracionalidad será abundantemente recompensada».[21]

* ¿Podemos encajar eso en la teoría estándar de la decisión diciendo que una consecuencia constante del hecho de que la persona realice la acción dominante en la situación de dilema del prisionero es que acabará entendiéndose a sí misma como una persona no cooperativa, y representando eso en la matriz del juego con una adición negativa, una adición de utilidad negativa, a lo largo de la entera columna de esa acción? Obsérvese que este componente de la utilidad sería una función de esa actitud hacia la acción de acuerdo con la situación de ésta en la estructura del conjunto de la matriz.

Tengo para mí que quienes así piensan conceden gran utilidad —¿se trata de utilidad simbólica?— a ser racional de acuerdo con su mejor estimación presente de cuáles son exactamente los principios que esto entraña. (Una sutileza adicional es distinguir entre quien confiere poco sólo a *un* principio particular, *UCE*, por ejemplo, y quien confiere algún peso a *UCE* y también un peso menor a *UEE*, pero concede también una gran utilidad simbólica —un gran peso— a seguir su mejor estimación *particular* de lo que implica la racionalidad.) Uno sospecharía que surgirían nuevas complicaciones si el hecho mismo de secundar un particular principio de decisión tuviera utilidad simbólica, o si la tuviera el comprometerse con un tipo particular de proceso de decisión.

Lo dicho acerca de la utilidad simbólica es lo mismo que decir que nuestras respuestas al dilema del prisionero se rigen, en parte, por nuestra concepción del tipo de persona que deseamos ser y por el tipo de maneras en que deseamos relacionarnos con otros. Lo que hacemos en una particular situación de dilema del prisionero implicará todo esto y lo evocará en distintos grados, según las precisas (*ratios* de diferencias entre) entradas de utilidad en la matriz y según las particulares circunstancias fácticas que dan origen a la matriz, circunstancias en las que una acción puede llegar a adquirir sus propios significados simbólicos, no simplemente a través de la estructura de la matriz.

Ya sabíamos todo eso, obvio es decirlo, al menos en tanto que tesis psicológica sobre por qué difiere la gente en sus respuestas a situaciones de dilema del prisionero. Sin embargo, el principio *VD* abre espacio para concepciones generales acerca de qué tipo de persona ser, en la medida en que esto se relaciona con y agrupa elecciones particulares, no simplemente como una posible *explicación* psicológica de por qué alguna gente se desvía de la racionalidad, sino como un componente legítimo, la utilidad simbólica, incorporado al procedimiento *racional* de decisión de las personas.

En un artículo seminal sobre el dilema del prisionero iterado, D.P. Kreps, P. Milgrom, J. Roberts y R. Wilson mostraron que el hecho de que ustedes asignen una pequeña probabilidad a mi realización de la acción cooperativa, o el de conferirla a mi creencia de que ustedes realizarán la acción cooperativa (o el de que asignen una probabilidad pequeña a mi creencia de que ustedes creen que yo realizaré una acción cooperativa) puede bastar para que sea racional para ustedes empezar realizando la acción cooperativa con objeto de alentarme a mí en mis acciones cooperativas o en mis creencias acordes con esas acciones.[22] Si ustedes creen que yo podría emprender

la acción cooperativa (o secundar un toma-y-daca) y creen que yo continuaré así sólo si *ustedes* se comportan de una determinada manera, entonces ustedes tendrán razones para comportarse como yo creo que podrían comportarse con objeto de alentar mi acción cooperativa.[23] Si la situación es común, ambos acabaremos (en determinadas circunstancias) realizando la acción cooperativa. Ello es que la estructura *VD*, cuando es de conocimiento común que ambas partes la secundan, confiere —tal es la promesa— cierta probabilidad a la creencia, por parte de cada parte, de que la otra creerá que la primera realizará la acción cooperativa, y por lo tanto, conferirá cierta probabilidad a que cada una de las partes realice la acción cooperativa. (Obsérvese que lo antedicho, y lo que se añadirá hasta el final del párrafo, *no* depende de la estructura *VD* completa, la cual incluye un peso para la utilidad simbólica. Basta la estructura más restringida antes expuesta, que se limita a ponderar la *UCE* y la *UEE*.) Y eso resulta no de una perturbación desviada de la plena racionalidad, no del ajuste de una de las partes al extravío irracional de la otra (o a la creencia de la otra de que ustedes podrían derivar en un extravío irracional), sino del conocimiento común de que todas las partes son *totalmente* racionales. Pues si el principio de maximización del valor decisional es un principio racional, normativamente deseable, entonces, si (como parece) el conocimiento común de la maximización-*VD* confiere cierta probabilidad a la realización de la acción cooperativa por parte de cada participante, el argumento de Kreps, Milgrom, Roberts y Wilson tiene validez incluso en condiciones de conocimiento común de que las partes son plenamente racionales.[24]

Lo interesante sería disponer de un resultado más preciso que la mera afirmación de que la acción cooperativa será realizada si las utilidades causales, evidenciales y simbólicas interactúan de forma oportuna. ¿En qué condiciones, con qué precisos pesos dentro de la estructura *VD* de uno de los participantes (o de ambos) elegirá una persona realizar la acción cooperativa en la situación de dilema del prisionero, o secundará la estrategia de toma-y-daca en el dilema del prisionero iterado?[25]

Hemos de limitarnos aquí a dar algunos pasos tentativos en punto a establecer una lista de supuestos apropiados de los que derivar resultados. Además de exigir que ambos jugadores secunden el principio *VD*, podemos añadir una forma extremadamente débil del supuesto de que cada uno debería esperar que el otro jugador se comportara como él mismo se comporta, se nutriera del componente *UEE*. Esta forma débil del principio predictivo dice que la probabi-

lidad evidencial condicional de que el otro jugador realice la acción *C'*, condicionada a que el primero realice *C*, es mayor que la probabilidad evidencial incondicional de que el segundo realice *C'*; y análogamente en lo que hace a su acción *D'* condicionada a la acción *D* del primer jugador. Un principio algo más fuerte, pero que aún se queda corto respecto del supuesto de simetría —según el cual el otro jugador racional actuará exactamente como ustedes—, sostendría que esas probabilidades evidenciales condicionales son, en la primera jugada, mayores que 1/2. Otro principio establecería que la persona confiere *alguna* utilidad simbólica (y algún peso simbólico a eso) a la realización de la acción cooperativa en la situación de dilema del prisionero. Por lo demás, podríamos presumir que la realización de la acción dominante *D* tiene en sí misma una utilidad simbólica negativa, que vendría a añadirse a la ausencia de la utilidad simbólica positiva de la cooperación.[26] Sea *S(A/B)* la utilidad simbólica de la acción *A*, dada la acción *B* de la otra persona. Si la persona I asigna utilidad simbólica positiva a la realización de la acción cooperativa, entonces *S(C/C')* es mayor o igual que *S(C/D')*, y cada una de ellas es mayor que (la magnitud negativa) *S(D/D')*, que a su vez es mayor que (la más negativa) *S(D/C')*. Cuando la estructura del dilema del prisionero se repite muchas veces entre las mismas dos personas, las ulteriores posibilidades de la cooperación mutuamente beneficiosa afectan a las utilidades del juego presente, incluido el primero. Además, la utilidad simbólica de una acción cambiará de un juego a otro, según hayan sido las acciones de la otra parte en el pasado. Podríamos entender que la utilidad simbólica de realizar la acción cooperativa desciende cuanto mayor uso hace la otra parte de la acción dominante, y que acaso desciende proporcionalmente al cociente del número de veces que la otra parte ha hecho uso de su acción dominante y del número de veces que ha hecho uso de su acción dominada. Cooperar con *esta otra parte* es tanto menos un símbolo de ser una persona cooperativa, cuanto más ha rechazado ella la cooperación. Por otro lado, cuanto más coopera la otra persona, tanto mayor será la utilidad simbólica de la acción cooperativa que ustedes realizan. Y una condición análoga vige aquí para la utilidad simbólica negativa de realizar la acción dominante. Esta desutilidad desciende también en términos absolutos cuanto mayor es el uso que hace la otra persona de su acción dominante, y se incrementa en términos absolutos cuanto mayor uso hace de su acción cooperativa. Hay que esperar que esas condiciones, junto con otros supuestos plausibles, lleven a resultados más precisos.

DISTINCIONES MÁS REFINADAS: CONSECUENCIAS Y FINES

Hemos examinado tres modos distintos de conexión entre la acción y los resultados —causal, evidencial y simbólica— y hemos sugerido que la teoría de la decisión ha de utilizar y reconocer explícitamente los tres modos. ¿Necesita también la teoría de la decisión proceder a distinciones más refinadas *dentro* de esas categorías? Por ejemplo, algunos especialistas en ética han sostenido que diferentes tipos de conexiones causales traen consigo diferentes pesos en situaciones de elección, aun cuando las probabilidades resultantes fueran idénticas. Hay una diferencia significativa —sostienen— entre dar origen a algo y meramente permitir que suceda o abstenerse de prevenirlo. (Y, por nuestra parte, podríamos considerar tipos adicionales de relación causal, como facilitar o ayudar a que suceda un acontecimiento.) Y algunos autores han formulado una doctrina del «efecto doble», sosteniendo que puede haber una diferencia moral (que puede a veces bastar para decidir si una acción es permisible) entre dar origen a algo, resultando ese algo de perseguirlo como un fin o como un medio para un fin, y dar origen a ese algo sabiendo que es un producto lateral de la búsqueda de algún otro fin. Es verdad que éstos son asuntos que se prestan a cierta controversia,[27] pero resulta sorprendente que la teoría de la decisión no se haya percatado hasta ahora de estas distinciones, verosímilmente importantes, y se conforme con una indistinta noción de «influencia causal». ¿Debería la teoría normativa de la decisión abrir espacio para estas distinciones y asignarles un papel, ya en su teoría de la elección en primera persona, ya en sus instrucciones para uso de expertos? Un lugar natural en el que podrían insertarse estas distinciones es la noción de utilidad condicional. Antes, al hablar de la teoría de la subasta, observamos que la teoría de la decisión debería hablar de *u*(resultado *O*/*A* se hace y *A* causa o consigue dar origen a *O*). El tipo preciso de vínculo causal que se da, en la segunda parte de esta condición, entre la acción y el resultado podría afectar a la utilidad del resultado conseguido *O*, esto es, arrojar distintas utilidades condicionales para *O*, y así, producir a veces distintas decisiones sin salirse de un principio que utiliza esas utilidades condicionales. ¿O es acaso el importe de esas distinciones enteramente simbólico, de manera que la incorporación de la utilidad simbólica a nuestra teoría bastaría para hacerles sitio?*

* ¿U ocurre más bien que esas distinciones son efectos contextuales, en el sentido de Tversky y Kahneman, que revelan las variaciones a lo largo (de las descripciones) de situaciones en las que debería haber invariancia? Véase Amos Tversky y Da-

Sugiero que veamos esas distinciones no como dicotomías, sino como elementos alineados a lo largo de una dimensión (no necesariamente continua). En realidad, no tenemos una dimensión, sino dos. La primera tiene que ver con la *importancia* del papel causal de la acción en relación con el efecto o el resultado o el estado de cosas resultante. Tenemos aquí (por lo menos) siete relaciones en que puede hallarse la acción con un estado de cosas. En orden decreciente de importancia, la acción puede: (1) causar el estado de cosas; (2) ayudar a o facilitar su ocurrencia (3) eliminar un obstáculo a su ocurrencia; (4) permitir su ocurrencia; (5) no prevenir y no evitar su ocurrencia (cuando alguna acción accesible a ustedes podría hacerlo); (6) no ayudar a o no facilitar su no-ocurrencia (cuando alguna acción accesible a ustedes podría hacerlo); (7) no ayudar a o no facilitar su no-ocurrencia (y *ninguna* acción accesible a ustedes podría hacerlo).

La segunda dimensión tiene también que ver con el papel causal de la acción en relación con el efecto o el resultado o el estado de cosas resultante, pero esta relación señala no la importancia de una acción, sino el *vigor* de la relación con ese resultado. La idea es que cuando se persigue algo como fin, valen ciertos subjuntivos para la persona. Ésta reorganizaría su conducta con objeto de alcanzar el fin (o de tener mejores ocasiones de alcanzarlo). En circunstancias ligeramente diferentes en las que *esa* acción no lograra alcanzar el fin, la persona haría alguna otra cosa que *alcanzaría* el fin; tendería a excluir las acciones que no tuvieran posibilidades de alcanzar el fin. Por otro lado, cuando algo es meramente un efecto lateral conocido de la acción, la persona no alteraría su conducta si llegara a darse el caso de que su conducta (actual o planeada) no produjera ese efecto lateral. Evidentemente lo haría si esa conducta produjera otro efecto lateral importante que ella deseara evitar. Es una cuestión del *espectro* de situaciones en las que la conducta se modificaría. Perseguir algo como un fin entraña subjuntivos que cruzan un espectro de circunstancias más amplio que el de actuar sabiendo que algo resultará como efecto lateral de la persecución de otro fin.

---

niel Kahneman, «Judgment under Uncertainty: Heuristics and Biases», *Science* 185 (1974); 1124-1131; reproducido por *Judgment under Uncertainty*, comp. Daniel Kahneman, Paul Slovic y Amos Tversky (Cambridge: Cambridge Univ. Press, 1982). ¿No parece sospechosa la relación entre la distinción dar origen/permitir que suceda y una línea de base, como lo es la relación entre la distinción ganancia/pérdida y *su* línea de base? Este último es, obvio es decirlo, el ejemplo predilecto de los efectos contextuales.

Entre estos dos casos cae el buscar algo solamente como medio para la realización de algún otro fin. Aquí la conducta sería reorganizada en algunas situaciones para conseguir los medios —a diferencia del caso de los efectos laterales—, pero cruzando un espectro de situaciones posibles más reducido que cuando el efecto es él mismo un fin. (Considérese, por ejemplo, esta situación posible en la que este efecto particular ya no sirve como un medio para el fin.)

En esa dimensión del vigor del papel causal, podemos distinguir (al menos) seis conexiones de una persona y de una acción con un efecto o resultado. La acción puede: (1) buscar el efecto como un fin; (2) buscar el efecto solamente como un fin. O puede (3) no buscar en absoluto el efecto. Y entre las acciones que no buscan el efecto, la persona podría: (3') conocer el efecto (que no se busca); (4') no conocer el efecto (que no se busca) que debería conocer; (5') no conocer el efecto (que no se busca), y no tener por qué conocerlo. O (6') el estado de cosas (que no se busca) ocurre por accidente.

Usando esas dos dimensiones, y sus respectivas categorías, podemos construir una matriz 7 × 6. (Si las dos dimensiones no son completamente independientes, algunas de las casillas pueden ser imposibles.) La relación de una persona y de una acción con su efecto (o con el estado resultante) se determinará por su localización en la matriz, esto es, por su posición a lo largo de las dos dimensiones.[28] ¿Debería tomar en cuenta la teoría de la decisión estas distinciones más refinadas relativas al modo de conexión causal entre una acción y un resultado? Y si es así, ¿cómo? ¿Hay también distinciones más refinadas *en el seno* de las conexiones evidenciales y simbólicas, distinciones que la teoría de la decisión debería acotar y tomar en cuenta? (¿No estarían acaso las categorías referidas al conocimiento, clasificadas dentro de la dimensión de vigor, mejor ubicadas en una dimensión evidencial?) No planteo estas cuestiones para responderlas aquí, sino para incluirlas en programa.

Los asuntos estudiados en estos dos capítulos sobre los principios y sobre el significado simbólico valen también para los principios éticos. Cuando agrupamos de un modo adecuado las acciones en una clase, una acción llega a valer por todas, y el peso de todas se hace descansar sobre cada una, confiriéndoles una desutilidad (simbólica) coordinada. Las restricciones deontológicas podrían exhibir ese mismo fenómeno. Al agrupar las acciones en un principio que las prohíbe —«no matarás»—, se inhibe el cálculo utilitario (o egoísta) separado de los costes y beneficios de *una* acción. La acción llega a valer por el grupo entero, y el peso del grupo descansa sobre ella. No es necesario que esto ocurra de modo tal que la res-

tricción sea absoluta, impidiendo la acción pase lo que pase, pero esa restricción constituye una barrera formidable, al inyectar en cualquier cálculo la acrecida desutilidad (simbólica) que trae consigo.[29]

Vale la pena recordar ahora la discusión realizada en el primer capítulo del significado simbólico de secundar principios éticos. La acción ética puede simbolizar (y expresar) el ser una criatura racional que se da leyes a sí misma, ser un miembro legislador del reino de los fines, ser fuente y reconocedor, en pie de igualdad con los demás, del valor y de la personalidad, etc. La utilidad de esas grandes cosas, simbólicamente expresada y ejemplificada por la acción, llega a incorporarse a la utilidad simbólica de esa acción, y así, al valor decisional de esa acción. De manera que estos significados simbólicos se convierten en parte de las razones de uno para actuar éticamente. Una persona que maximiza la utilidad —en sentido amplio— de una acción, esto es, que maximiza su valor decisional (*VD*), puede verse inducido a realizar acciones éticas. Tal persona perseguiría sus *propios* fines (que no tienen por qué ser egoístas). Por emplear las categorías de Amartya Sen, esa persona se comprometería más bien con la persecución de auto-fines, no con la actividad de *no* perseguir marginalmente *su* propio fin individual global.[30] Obsérvese, no obstante, que, si caer en esta categoría de no perseguir marginalmente su propio fin individual global llegara a tener utilidad simbólica para ella, entonces ésta entraría también en su *VD*. Llegada a este punto, cuando la persona actúa tomando en cuenta esta utilidad simbólica, ¿persigue de nuevo su propio fin, esto es, *su VD* revisado, de manera que su intento (dentro de la estructura *VD*) de entrar en la otra categoría de Sen está condenado al fracaso? Cualquiera que sea la decisión que tomemos, valdrá la tesis más general que hemos establecido. Ser éticos es nuestro modo más efectivo de simbolizar (una conexión con) lo que más valoramos, y eso es algo de lo que ninguna persona *racional* querría prescindir.

En el primer capítulo tuvimos ocasión de analizar varias funciones de los principios. Aceptar y adherirse a un particular principio, dijimos, podría considerarse una acción (general) *A* y tratarse en un marco de la teoría de la decisión, que era, por mucho, un marco instrumental. Ahora hemos presentado un marco alternativo para la teoría de la decisión, un marco que incluye aspectos evidenciales y simbólicos, no simplemente la instrumentalidad causal. En este marco, una acción de aceptación de un principio tendrá un valor decisional *VD*, y será elegida (entre varias alternativas) cuando tiene

un valor decisional máximo. Este marco más amplio despeja el camino para revisar la discusión de por qué hemos de tener principio alguno en general —por qué, si tenemos este principio *VD*, tendremos algunos más— y de por qué tenemos algunos en particular.

CAPÍTULO 3

LA CREENCIA RACIONAL

¿Cuándo es racional una creencia? ¿Por qué queremos que nuestras creencias sean racionales, cómo podemos decir si lo son y cómo podemos mejorar su racionalidad?

Dos temas permean la producción filosófica. En primer lugar, que la racionalidad es una cuestión de razones. La racionalidad de una creencia depende de las razones para sostener tal creencia. Éstas pueden ser razones para pensar que la creencia es verdadera (o acaso para pensar que tiene alguna otra virtud cognitiva deseable, como fuera explicativa). En segundo lugar, que la racionalidad es una cuestión de fiabilidad. ¿Conduce el proceso o el procedimiento que produce (y mantiene) la creencia a un alto porcentaje de creencias verdaderas? Una creencia racional es una creencia que surge de algún proceso en cuya producción de creencias verdaderas (o investidas de alguna otra virtud cognitiva deseable) confiamos.

Ninguno de los dos temas por sí solo agota nuestra noción de racionalidad. Las razones sin fiabilidad parecen hueras; la fiabilidad sin razones, ciega. Juntas, constituyen una poderosa unidad, mas ¿cómo se relacionan exactamente entre sí, y por qué?

Resulta natural pensar en la racionalidad como en un proceso orientado a un fin. (Esto vale tanto para la racionalidad de la acción, como para la racionalidad de la creencia.) El estereotipo de la conducta en las sociedades tradicionales es la gente que actúa de determinada manera porque las cosas siempre se han hecho así. En cambio, la conducta racional trata de lograr objetivos, deseos y fines que la gente tiene. En esa concepción instrumental, la racionalidad consiste en el logro efectivo y eficiente de objetivos, fines y deseos. Respecto de los objetivos mismos, una concepción instrumental tiene poco que decir.* Si son procedimientos racionales aquellos que

* «"Razón" tiene un significado perfectamente claro y preciso. Significa la elección de los medios adecuados a un fin que ustedes desean conseguir. No tiene nada que ver con la elección de fines.» Bertrand Russell, *Human Society in Ethics and Politics* (Londres: Allen and Unwin, 1954) [trad. cast.: *Sociedad humana*, Madrid, Cátedra, [3]1987], pág. VIII. «La razón es puramente instrumental. No puede decirnos adónde ir; a lo sumo, puede decirnos cómo ir. Es un arma mercenaria que puede em-

fiablemente logran objetivos definidos, una acción es racional cuando es producida por tal procedimiento, y una persona es racional cuando emplea adecuadamente procedimientos racionales.

Esa noción de la racionalidad de una acción se refiere al proceso o procedimiento del que surge. Sin embargo, la noción estándar de la racionalidad de una acción que nos presenta la teoría de la decisión se refiere sólo a que esa acción maximiza la utilidad esperada. Pero una acción podría lograr objetivos o maximizar la utilidad esperada sin que se hubiera llegado a ella racionalmente. Podría haberse tropezado con ella por accidente, o haberse hecho inadvertidamente, o resultar de una serie de errores de cálculo que se cancelaran entre sí. Tal acción, entonces, habría sido lo mejor que podía hacerse (dados los objetivos de esta persona), pero no habría sido hecha racionalmente.

Parecería, pues, que la teoría de la decisión está obligada a referirse también al proceso o procedimiento generador de una acción si quiere ser una teoría de la racionalidad. La fórmula de la utilidad esperada (o del *VD*) distingue a una acción como la mejor. Una acción es racional cuando está generada por un proceso que tiende a producir las mejores acciones, aquellas con máximo *VD*. (Para decirlo todo, en la evaluación de la optimalidad de ese proceso entran también consideraciones y criterios procedentes de la teoría de la decisión.) La teoría de la decisión por sí misma es una teoría de la mejor acción, no de la acción racional. Cuando pensamos en aplicar la teoría de la decisión, tendemos a ignorarlo. Pues, una acción así propiciada ¿no está acaso generada por un proceso que arroja fiablemente las mejores acciones? ¿Y no la hace eso racional? Esto es, no obstante, una tesis *empírica*. La teoría de la decisión perfila la mejor acción, pero nosotros podríamos ser usuarios poco fiables de esa teoría: podríamos tender a ignorar determinados factores, incurrir en errores de cálculo, o cometer algún otro yerro en la aplicación de la teoría. Podría darse el caso de que otro procedimiento arrojara un porcentaje mucho mayor de acciones óptimas, aun cuando el *criterio* de optimalidad lo siguiera suministrando la teoría de la decisión.[1]

También la racionalidad de la creencia puede entenderse en esos términos instrumentales. En este caso, los particulares objetivos a conseguir están bien definidos: la verdad, la elusión del error, la fuer-

---

plearse al servicio de los objetivos que tengamos, buenos o malos.» Herbert Simon, *Reason in Human Affairs* (Stanford: Stanford Univ. Press, 1983), págs. 7-8.

za explicativa, etc. Supongamos por un momento que hubiera un solo objetivo cognitivo: creer la verdad. Una creencia racional, entonces, sería una creencia a la que se llegaría mediante un proceso en cuya producción de creencias verdaderas pudiera confiarse. Tal proceso incluiría no sólo la adquisición de una nueva creencia, sino también la forma en que una creencia existente se mantiene, se elimina o se revisa.[2] En general, sin embargo, me limitaré a estudiar el problema de la adquisición.

Se plantean cuestiones acerca del contexto en el que se supone que se da la fiabilidad. ¿Tiene un procedimiento que ser fiable sólo cuando se usa, o tendría que serlo siempre que pudiera usarse en el mundo tal como es, o serlo en mundos similares al mundo real, o serlo en mundos que en general fueran como creemos que es el mundo (sea o no correcta esa creencia)?[3] También se plantean cuestiones acerca del grado de fiabilidad que debe tener un procedimiento racional. ¿Tiene que ser el procedimiento más fiable de los disponibles, o acaso puede ser racional una creencia si viene generada por un procedimiento bastante fiable que, sin embargo, no es el mejor? ¿Debe la confianza en un procedimiento ser mayor que la desconfianza de él, o es por ventura posible que una creencia acerca de la explicación de un fenómeno sea racional cuando surge del proceso más fiable para llegar a explicaciones correctas, aun cuando esa fiabilidad sea inferior al 50 por ciento? Y, al evaluar un procedimiento, ¿no deberíamos dejar de limitarnos a observar el porcentaje de veces que el procedimiento da resultados correctos, para observar también lo que sucede cuando el procedimiento da malos resultados y para estimar cuán malos son esos resultados? Así como la acción racional no es siempre la que arroja la probabilidad más alta de un resultado deseado, así también el procedimiento racional que hay que usar no siempre es el que arroja la probabilidad más alta de conseguir el objetivo. Vienen aquí a cuento las consideraciones de la teoría de la decisión.*

* La fuente moderna de las concepciones de la fiabilidad es Charles Peirce, quien habló de la validez de las reglas de inferencia («principios rectores») expresadas en términos de porcentaje de veces que, cuando las premisas son verdaderas, sus conclusiones son verdaderas. Las reglas deductivamente válidas logran el 100 por ciento, mientras que las reglas inductivas válidas conseguirían un porcentaje muy elevado. Véase Charles S. Peirce, «The Fixation of Belief», en sus *Collected Papers* (Cambridge, Mass.: Harvard Univ. Press, 1931-1958), 5: 223-247, reproducido en *The Philosophy of Peirce: Selected Writtings*, comp. J. Buchler (Nueva York: Harcourt, Brace, 1950), págs. 5-22. Obsérvese que un elevado porcentaje, en el sentido de Peirce, no es suficiente para justificar la *aplicación* de la regla de inferencia. Supongamos que, normalmente, cuando un enunciado del tipo *p* es verdadero, un enunciado

Parece que los dos temas, razones y fiabilidad, son de fácil conexión: ustedes consiguen una creencia verdadera con mayor frecuencia si ustedes sostienen esa creencia con razones que le sirven de apoyo. Logramos que los procesos que generan nuestras creencias sean fiables confiriendo a las razones el papel más destacado.

---

correspondiente del tipo *q* es verdadero también. (Podríamos formularlo como un enunciado estadístico sobre porcentajes, o con probabilidades.) ¿Significa eso que puedo inferir fiablemente *q* de *p*, y contar con que estoy en lo cierto (aproximadamente) este tanto por ciento de veces? No. Podría darse el caso de que cuando yo creo un enunciado del tipo *p*, el correspondiente enunciado del tipo *q* sea normalmente *no* verdadero. Después de todo, las ocasiones que se me ofrecen para creer un enunciado de tipo *p* no son una muestra aleatoria de las situaciones en las que *p* es verdadero, y puede que no sean una muestra representativa, quizá debido a un sesgo de la evidencia que fluye hacia mí, o debido a un sesgo ínsito en mí. Incluso cuando hay un enunciado estadístico verdadero de la forma: «Normalmente, cuando yo *creo* en un enunciado del tipo *p*, un enunciado del tipo *q* es verdadero», incluso en ese caso, cuando yo infiero *q* del *p* en el que creo, *q* normalmente es falso. Pues las veces en que yo infiero *q* de *p* (o realizo cualquier inferencia partiendo de *p*) no tienen por qué constituir una muestra aleatoria, o representativa, de las veces en que yo creo *p*. ¿Tendríamos entonces que decir más bien: «Estadísticamente, normalmente, cuando yo creo *p* e infiero *q* de *p*, entonces *q* es verdadero»? ¿Es este enunciado lo que nos autoriza a inferir *q*? Mas, inferir ¿de acuerdo con *qué* regla de inferencia? ¿Necesitamos determinar, dentro de la regla misma, que *éste* es un caso representativo en el que el tipo *p* es verdadero, se cree en él y se usa como base de inferencia de acuerdo con *esta* regla? Con una inferencia fundada en un principio rector universalmente verdadero no tenemos por qué preocuparnos de la particular ocasión en que se realiza nuestra inferencia. Pero con una inferencia fundada en un principio estadístico, tenemos que preocuparnos de si *esta* ocasión en que se realiza la inferencia es una ocasión representativa.

La idea en que arraiga un análisis de fiabilidad de la racionalidad es que la racionalidad se aplica, por lo tanto, a un proceso o procedimiento, y derivativamente, a una inferencia, a una creencia o a una acción particulares como un caso de tal procedimiento. Ese caso hereda su racionalidad del procedimiento que ejemplifica. Cuando el rasgo deseable del procedimiento es su fiabilidad, el que arroja un alto porcentaje de creencias verdaderas, entonces el procedimiento arrojará probablemente una creencia verdadera. ¿Nos dice eso, de una particular creencia generada por el proceso, que es probable que *ella* sea verdadera? Esta creencia o inferencia puede caer bajo muchos procedimientos posibles, y cae también bajo muchas otras clasificaciones. Inferir que tiene una determinada probabilidad de ser verdadera partiendo de su pertenencia a una clase —en este caso, a la clase de creencias generadas por un particular procedimiento—, dada una información probabilística general sobre esa clase, es realizar (lo que en teoría de la probabilidad se llama) una *inferencia directa*. Puesto que los casos particulares pueden caer bajo muchas clases de referencias distintas, resulta delicado formular criterios de validez para una inferencia directa que genere un juicio probabilístico (separable) sobre un caso particular. Véase C.G. Hempel, «Inductive Inconsistencies», en su *Aspects of Scientific Explanation* (Nueva York: Free Press, 1965) [trad. cast.: *La explicación científica. Es-*

Mas, ¿qué es lo que hace que algo sea una razón, y por qué fundar una creencia en razones contribuye a sostener creencias verdaderas? ¿Por qué el objetivo de la creencia debería ser la verdad? ¿Sería alguien irracional por el hecho de que sus creencias fueran generadas por procesos que sirvieran (fiablemente) a objetivos distintos, tales como hacerle feliz o bienquisto? ¿Podemos por ventura hallar reglas o condiciones precisas para definir qué creencias son racionales y cuáles son irracionales?

OBJETIVOS COGNITIVOS

¿Cuándo es racional una creencia porque ha sido generada (y mantenida) por un procedimiento que fiable y eficientemente consigue ciertos objetivos? ¿Cuáles son esos objetivos? Suele decirse que esos objetivos son objetivos *cognitivos* —creer la verdad, evitar el error, o quizá un mixto más amplio, que incluiría la fuerza explicativa, la contrastabilidad y la fertilidad teórica—.[4] La noción misma de objetivo cognitivo no está tan bien delimitada. La verdad, sí, y también la fuerza explicativa, la fertilidad teórica y el alcance. La simplicidad, empero, ¿es un objetivo cognitivo? ¿Y la facilidad de computación? La intuición mística de la naturaleza de Dios, ¿es un objetivo personal cognitivo o no cognitivo? ¿Y cuál de estas dos cosas es la iluminación, en el sentido de las teorías orientales?

El objetivo cognitivo primario discutido por los filósofos es la verdad. ¿Por qué es la verdad un objetivo? Una respuesta es que la verdad, o la creencia en la verdad, es intrínsecamente valiosa. ¿Toda la verdad? A mí no me inquieta en absoluto la idea de que albergo algunas creencias falsas acerca de algunos asuntos —las capitales de algunos estados, pongamos por caso—, y hay muchas verdades de cuyo conocimiento no me preocupo en absoluto —las referidas al número exacto de granos de arena que hay en cada playa del mun-

---

tudios sobre la filosofía de la ciencia, Barcelona, Paidós, 1988], págs. 53-79; Henry Kyburg, «Randomness and the Right Reference Class», *Journal of Philosophy* 74 (1977): 501-521; Isaac Levi, *The Enterprise of Knowledge* (Cambridge, Mass.: M.I.T. Press, 1980), capts. 12, 16. En la noción en términos de fiabilidad de la racionalidad de la creencia o de la inferencia, esos asuntos delicados no afectan sólo a la formulación de los principios correctos para un tipo particular de inferencia; la infección llega al mismísimo concepto de racionalidad. Pues la racionalidad de los casos particulares se *define* en términos de una inferencia directa que parte de información probabilista sobre una clase.

do—. No cualquier hecho merece ser conocido, ni merece el que se genere una creencia verdadera sobre él, aunque para determinados propósitos podría ser interesante llegar a tener ese conocimiento. El conocimiento de algunas cosas *es* valioso por sí mismo, según creo —por ejemplo, verdades profundas que explican una amplia gama de hechos, *la* gran teoría explicativa, si es que hay alguna, la verdad acerca de cómo se originó el universo—. Ciertamente, podemos llegar a desarrollar curiosidad intelectual acerca de una gama particular de hechos.

Parece razonable pensar que nuestro interés originario en la verdad tuvo un fundamento instrumental. Las verdades nos servían mejor que las falsedades, y mejor que la falta completa de creencias, a la hora de lidiar con los peligros y las oportunidades del mundo.* No era necesaria la verdad perfectamente exacta, sólo una creencia que fuera *suficientemente verdadera* para dar (más) resultados deseables cuando actuáramos de acuerdo con ella. Lo necesario y deseable eran «verdades serviciales»; y para que sea servicial no es necesario que una creencia sea exactamente verdadera.** La verdad, entonces, se parecería más bien a lo que John Rawls ha llamado un bien primario, algo que resulta útil para una *muy* amplia gama de propósitos —casi todos— y que, por lo mismo, resulta deseable y trae consigo beneficios (casi) independientemente de cuáles sean nuestros propósitos particulares.[5] De manera que podríamos desear creencias verdaderas y llegar a interesarnos por la verdad porque las creencias verdaderas son útiles para una muy amplia gama de propósito. Sin embargo, eso confinaría nuestro interés por la verdad al ámbito de lo instrumental —al menos, al comienzo—.[6]

¿Tenía, pues, razón William James cuando decía que la verdad es lo que funciona? Podríamos interpretar la afirmación de James

* La base instrumental de nuestro interés por la verdad no es que la gente desee creer verdades porque reconoce que esto es instrumentalmente útil. Es más bien ésta: a causa de la utilidad de creer en verdades aproximadas, se produjo una selección evolucionaria de algún tipo de interés por la verdad de la creencia, como por ejemplo cierta curiosidad por descubrir verdades. Aunque no habría que excluir tampoco el desarrollo de algún deseo instrumental explícito de la verdad.

** En contra de la idea, según la cual creemos *p* y sólo nos preocupamos de si esa creencia fue y podría seguir siendo servicial, alguien podría argüir que lo que creíamos era: «Es aproximadamente verdad que *p*», y lo que queremos es que sea *esto* lo que es exactamente verdadero. Mas no cualquier aproximación resultaría servicial en cualquier contexto. De modo que acaso lo que creíamos era «*p* es servicial», y era esto lo que deseábamos que fuera exactamente verdadero. Con todo, ¿por qué pensar que esto muestra que nos preocupábamos por la verdad más que por la servicialidad?

como referida al *valor* de la verdad, no a su naturaleza. En vez de sostener que la verdad es simplemente «servicialidad», podríamos construir la verdad como la propiedad —cualquiera que sea— que subyace a y explica la servicialidad. Si una propiedad subyace a la servicialidad de varios enunciados sobre objetos diferentes, tal propiedad tendrá que estar muy general y abstractamente formulada. Las varias teorías de la verdad —correspondencia, coherencia, etc.— serían entonces hipótesis explicativas, conjeturas acerca de la naturaleza de la propiedad que subyace a y explica la servicialidad.*

Por lo general, creer en un enunciado falso disminuye la proporción de creencias verdaderas; pero bajo ciertas circunstancias la creencia en una falsedad podría ayudarnos a maximizar esa proporción. (Si él cree esta cosa que es falsa, yo podré contarle muchas verdades que, de otro modo, él no podría llegar a descubrir. O yo le regalo a él una matrícula para una universidad en la que aprenderá mucho. Si él cree, falsamente, que tuvo resultados excelentes en un determinado test, estará motivado para aprender con éxito muchos enunciados verdaderos.) ¿Es racional para él creer este enunciado falso, seguir un proceso que arroja esta creencia? Aun no siendo verdadero, un enunciado puede llegar a ser creído merced a un proceso que maximiza la proporción de creencias verdaderas. Si el objetivo cognitivo es maximizar la proporción de creencias verdaderas, entonces esta creencia falsa, derivada de un proceso que efectivamente sirve a ese objetivo, es racional. Pero si el propósito cognitivo de la verdad se manifiesta él mismo como una restricción aplicable a este caso particular —«no creas nada falso»—, entonces esta creencia falsa no sirve al propósito y es, por lo tanto, irracional.[7]

Esas dos formas no son las únicas que puede cobrar el propósito. Amartya Sen ha propuesto una estructura, dentro de la cual se confiere a los dos tipos de objetivos —a saber: que el agente no haga él mismo esta vez algo de tipo *T*, y que, a un tiempo, maximice la cantidad de no-*T* hecho por todos— un peso separado (como parte de un maximando que puede asimismo incluir otros objetivos).[8] En esa estructura, entre los objetivos cognitivos de un agente se incluirán tanto creer la verdad (evitar la creencia en una falsedad) esta vez, cuanto maximizar su proporción de creencias verdaderas. De modo que hay (al menos) tres formas distintas de perseguir el objetivo cognitivo de la verdad; como una restricción lateral, como un objetivo de maximización de la proporción de creencias verdaderas,

* ¿Y si al final la servicialidad de distintos tipos de enunciados se explicara por distintas propiedades?

o como el resultado de ponderar este último objetivo con el objetivo de evitar una creencia falsa esta vez. De aquí que no baste con decir que una creencia es racional cuando se llega a ella por un procedimiento que efectiva y eficientemente logra el propósito de la creencia verdadera. ¿Qué estructura ha de usarse para evaluar la racionalidad de un procedimiento? ¿O hay acaso tres nociones distintas y todas ellas legítimas de racionalidad, adecuada cada una para distintos fines?

Hay situaciones en las que parece claro que creer la verdad *no* servirá a otros objetivos importantes de una persona. Esas situaciones han sido objeto de investigación especializada bajo el marbete de «ética de la creencia». Por ejemplo, a una madre se le presenta evidencia judicial de que su hijo ha cometido un grave crimen, una evidencia que convence a todos los demás pero que, de ser creída por ella, tornaría su vida miserable.[9] ¿Es racional para ella creer que su hijo es culpable? Un análisis bayesiano podría mostrar que ella debería llegar a una conclusión diferente de la de los demás: ella sabe más acerca de su hijo que ellos, de manera que puede comenzar asignando una probabilidad previa diferente. Supongamos, empero, que su diferente probabilidad previa se basa simplemente en su amor por el hijo y en su mala disposición para creer que podría ser culpable de un crimen así. Todavía podría argüirse que es racional al creerle inocente, porque esa creencia maximiza su utilidad esperada una vez considerados todos sus objetivos —¿y no es éste el criterio en que se funda una acción racional?—.[10] Análogamente, una persona podría tomar en cuenta los probables efectos morales negativos que para sí misma acarrearía el creer ciertas proposiciones —por ejemplo, que hay diferencias heredables de inteligencia entre grupos raciales— y, con esa base, abstenerse de examinar determinada evidencia o de generar ciertas creencias. (Esos efectos negativos serían efectos no sometidos al control volitivo de la persona, o controlables sólo con grandes dificultades, o a un coste muy elevado.)

Podríamos distinguir (1) la proposición de que *p* es lo que racionalmente hay que creer de (2) creer que *p* es lo que racionalmente hay que hacer. La perspectiva de la utilidad esperada podría aplicarse a (2), considerando que creer *p* es una acción a disposición de la madre; pero no parece aplicable a (1), al menos respecto de *todos* los objetivos que la madre pudiera tener. Por lo que hace a (1) —la proposición de que *p* es lo que hay que creer racionalmente—, parecería que sólo deberían pesar las consideraciones evidenciales. Y aun si la perspectiva de las consideraciones evidenciales se revela-

ra instrumental, los objetivos relevantes deben ser únicamente objetivos *cognitivos* —y la felicidad de la madre, cualquiera que fuere su importancia, no es un objetivo cognitivo—.[11]

No obstante, si los mismos objetivos cognitivos tienen una base y una justificación enteramente instrumentales, ¿no sería racional «mirar al través» de los objetivos cognitivos, hacia los propósitos últimos a que supuestamente aquéllos sirven? Cuando resulta manifiesto que esos propósitos no están bien servidos, ¿no deberíamos ignorar el objetivo cognitivo para perseguir más directamente, los propósitos últimos? (Si tal es el caso, la proposición de que su hijo es inocente podría llegar a ser la creencia racional para la madre.) La noción de qué proposición es racional creer, la noción (1) antes mencionada, podría *definirse* entonces más bien en términos de objetivos *cognitivos*, pero, ¿no resultaría así arbitrariamente truncada la noción? Aun si fundados instrumentalmente, ¿llegan los objetivos cognitivos a adquirir una autoridad por sí mismos, incluso enfrentados a los propósitos últimos que les dieron origen como objetivos?[12]

Las concepciones holistas de la creencia sostienen que la adición de creencias particulares genera efectos de onda a través del cuerpo de creencias. Modifica muchas otras creencias, altera las probabilidades previas de muchas hipótesis que entran en futuros cálculos bayesianos, plantea problemas nuevos y recalcitrantes para la unificación explicativa global de las creencias de uno, etc. Sería poco avisado, pues, atender únicamente a los inmediatos efectos personales gratificantes de creer algo. Los efectos de largo alcance que tiene el introducir esta creencia serán imposibles de calcular, sobre todo si, para acomodarla, uno tiene que modificar también los propios procedimientos de formación de creencias. Incluso en el nivel personal, hay una muy robusta presunción de que los efectos inmediatamente beneficiosos serán cancelados por las consecuencias de onda de otras creencias falsas resultantes y por las ulteriores consecuencias de albergar estas últimas.

Con todo, en cualquiera de los casos concretos, los efectos personales de creer un determinado enunciado pueden ser claros, y puede resultar tentador el dejarse llevar sin más y creerlo, cualquiera que sea la evidencia. Para evitar este tipo de tentaciones fuertes y distinguidas, peligrosas si se cede a ellas con frecuencia, o aun una sola vez, una persona podría adoptar un *principio*, según el cual habría que creer sólo lo que es verdad, sólo lo que la evidencia muestra como (probablemente) verdadero. De acuerdo con nuestra noción del modo en que los principios le ayudan a uno a evitar tentaciones

cuando adopta ese principio, creer una falsedad en esta ocasión particular llega a valer por muchos casos de creencia en falsedades; creer la verdad en esta ocasión llega a valer por muchos casos de creencia en verdades. Creer una verdad particular llega a tener una utilidad simbólica desvinculada de *sus* consecuencias reales. Los objetivos cognitivos podrían llegar a adquirir una autoridad independiente aun en situaciones en las que las consecuencias *locales* de ignorarlos fueran más beneficiosas. Volveré a las cuestiones de la ética de la creencia más adelante, en la sección «Reglas de racionalidad», dentro de este mismo capítulo.

La sensibilidad a las razones

La racionalidad de una creencia puede derivar del proceso merced al cual se llega a la creencia y se la mantiene, mas no cualquier vía (concebiblemente) efectiva para llegar a una creencia serviría para distinguirla como una creencia racional. Si ser golpeado en la cabeza o ingerir mescalina fueran vías para llegar a tener creencias verdaderas sobre un determinado asunto —un asunto distinto de si uno ha sido golpeado en la cabeza o ha ingerido esa substancia—, entonces esta creencia misma podría no ser una creencia racional. (Para alguien que supiera eso, sin embargo, podría ser racional elegir que le golpearan en la cabeza con objeto de adquirir una creencia verdadera.) La racionalidad de una creencia está conectada con una densa red de razonamiento, inferencia y evaluación de evidencia en cadenas de enunciados que se solapan. La observación puede nutrir a esa red, pero a partir de cierto nivel de descripción el proceso es proposicional. Esto muestra que la racionalidad no es simplemente un tipo *cualquiera* de instrumentalidad. Requiere cierto tipo de instrumento, a saber: razones y razonamiento. Supongamos, pues, que un particular procedimiento es una vía fiable para llegar a una creencia verdadera. Si una acción o una creencia arrojada por tal procedimiento ha de ser racional, no sólo tiene el procedimiento que entrañar una red de razones y razonamiento, sino que esto debe explicar (en parte) *por qué* el razonamiento es fiable. Las razones y el razonamiento contribuyen a la fiabilidad del procedimiento.[13]

La racionalidad no entraña solamente el hacer o el creer algo por las razones *en favor de* ello, sino tomar también en cuenta (algunas) razones *en contra de* ello. Karl Popper puso de relieve la importancia, para la ciencia, de buscar datos o evidencia contrarios a la hipótesis o teoría. Las confirmaciones, observaba, son fáciles de en-

contrar. Lo que distingue a una teoría científica es que excluye cierta evidencia o ciertos hechos, y sometemos la teoría a prueba escrutando aquellos ámbitos en los que más probablemente podría revelarse falsa, no acumulando más y más casos en los ámbitos en los que probablemente es verdadera.[14] Destacar eso resulta saludable aunque no compartamos el punto de vista de Popper, según el cual no hay razones *a favor de* (ni sostengamos que las únicas razones a favor de son los informes sobre el fracaso de la búsqueda de razones en contra de). La racionalidad implica tomar en cuenta las razones a favor *y* las razones en contra. La creencia o la acción no sólo debe estar causada (por la vía correcta) por las razones a favor y por las razones en contra; debe ser *sensible* a esas razones. Dentro de la latitud de algún espectro de variación en la naturaleza, o en la fuerza, o en el balance de esas razones, si las razones fueran diferentes, entonces la acción o la creencia sería también diferente.[15] La creencia o la acción debe ser positivamente sensible: si las razones son más robustas, la creencia no debe desaparecer o hacerse más débil; si las razones son más débiles, la creencia no puede robustecerse.

El artículo prototípico en una revista de filosofía está organizado para inducir en sus lectores creencias racionales. Normalmente, se avanza como digna de ser creída una proposición o una tesis filosófica, y se consideran las razones a favor y en contra. Entre las razones a favor de la tesis están: enunciados generales y aceptables de los que se sigue la tesis; otras cosas aceptables que casan con la tesis o se compadecen bien con ella; las consecuencias de la tesis que resultan aceptables, y así, le dan apoyo; casos que caen bajo la tesis o ejemplos que casan con ella, y le suministran así alguna evidencia. Entre las razones consideradas en contra de la tesis están: posibles objeciones a la misma (a las cuales se replica, se les quita fuerza, se las socava, o de un modo u otro se las sortea); posibles contraejemplos (a los cuales se les neutraliza, *o* se los usa para modificar la tesis y conseguir otra formulación que no sea objeto del contraejemplo, proponiéndose ahora el enunciado modificado como digno de ser creído racionalmente). Hay una razón en contra de un enunciado *p* que merece una atención especial, a saber: que una proposición alternativa *q*, la mejor o la más plausible alternativa a *p*, pueda ser más digna de ser creída racionalmente que *p*. La práctica consiste en plantear determinadas objeciones a *q*, dificultades o contraejemplos que se presumen suficientes para eliminarla o para mostrar por qué no debería ser aceptada. Raramente se dice que las objeciones a *q* no son más graves que las planteadas a *p*, pero que *p*

tiene mejores razones a su favor. Y más raramente aún se otorga a *q* el mismo tratamiento exhaustivo que a *p*. Aun así, todo esto monta tanto como una consideración de las razones a favor y en contra de la tesis, dejando al lector en una mejor posición para creerla racionalmente.

Quizá deberíamos empezar diciendo que la racionalidad entraña la sensibilidad respecto de factores relevantes, respecto de todos los factores y sólo respecto de los relevantes. Una tesis adicional es que los factores relevantes son *razones*. Y aun otra tesis más —una que acaso no valga para todos los dominios—: que esas razones se dividen netamente en dos categorías, las que son razones *a favor* y las que son razones *en contra*. ¿Pero exactamente de qué modo es una creencia racional sensible a todas las razones a favor y en contra, de qué modo está determinada la credibilidad de la creencia por las razones a favor y por las razones en contra? Sería demasiado simple decir que una persona sostiene creencias con objeto de maximizar el peso *neto* de las razones para sus creencias, esto es, la medida del peso de las razones a favor menos la medida del peso de las razones en contra de la creencia. Podría haber alguna interacción entre las razones a favor y las razones en contra. O el peso de las particulares razones aducidas a su favor podría depender de qué razones en contra se les oponen; incluso el que algo *sea* una razón a favor podría depender de qué razones haya en contra.

Además, el particular peso que tenga una razón puede depender de otros factores que no son ellos mismos razones a favor o en contra. Un enunciado *r* puede ser una razón para creer *S*, y sin embargo *q* puede socavar a *r* como tal razón, no porque constituya una razón para pensar que *r* mismo sea falsa, o que lo sea *S*, sino por tratarse de una razón para pensar que (en este contexto) *r* es una razón más débil para *S*, o no es ninguna razón en absoluto para *S*.[16] (Análogamente, podría haber cosas que incrementaran el peso de razones: agravantes.) Por eso el peso neto de las razones a favor no viene fijado solamente por razones; ese valor depende de los socavantes (y de los agravantes) que lo circundan.

Eso sugiere la formulación de un modelo de red neural, de razones a favor y en contra. Una razón *r* para el enunciado *S* manda, a través de un canal, una señal con un cierto peso positivo al nódulo *S*. Una razón *r'* en contra de *S* manda, a través de un canal, una señal a *S* con un cierto peso negativo. Un socavador de la razón *r* para *S* mandará, a través de un canal, una señal con un determinado peso para reducir el peso (quizá hasta cero) en el canal entre *r* y *S*. El resultado de esta red es un valor de credibilidad para el enunciado

S. (Seguiremos explorando esta estructura en la próxima sección.) En ese marco hay espacio para muchos tipos de razones con pesos diferentes. Así, podríamos esperar capturar muchas máximas metodológicas procedentes de la filosofía de la ciencia, ya dentro de la red, ya como fenómeno emergente del conjunto de la red.

De lo que la persona racional se preocupa, sin embargo, es de la verdad. Saca el balance neto de las razones de que dispone o de que sabe, y lo usa para estimar o predecir esa verdad. Acaso no sea razonable, empero, estimar la verdad mediante una extrapolación directa a partir del balance neto de las razones de que disponemos; pues esas razones quizá sean un indicador sesgado de esa verdad. El proceso que nos hace accesibles las razones acaso admita diferencialmente ciertos tipos de razones, aquellas que apuntan a la verdad del enunciado, pero no aquellas que apuntan a la falsedad del mismo. Seríamos entonces poco avisados si fundáramos nuestra creencia únicamente en las razones que tenemos, sin pararnos a considerar la representatividad de esas razones.

He aquí una imagen algo artificial, pero sugerente. Pensemos en las razones de una persona como en una muestra de la totalidad de las razones relevantes respecto del enunciado *S*. Esa totalidad puede incluir los hechos que otras personas conocen, los hechos que podrían llegar a ser averiguados, etc. (añadiendo acaso alguna restricción que excluya el enunciado *S* mismo, aun si otra persona lo conoce, así como algunos otros enunciados que implican *S*). La cuestión, entonces, es si las razones de esta persona constituyen una muestra sesgada o no representativa de la totalidad de las razones. La persona misma puede tener alguna razón adicional para pensar que ese sesgo existe.* No digo que la persona racional comenzará

* Pero si incluimos esta última razón entre las razones a favor y en contra, ¿podemos entonces proceder a una extrapolación en línea recta? Ello es que esta última razón acaso no afecte a *x* directamente como una razón a favor o en contra de alguna conclusión acerca de ella; puesto que afecta más bien a nuestros procesos de adquisición de información, es mejor tratarla a otro nivel.

Bernard Williams sostiene que las únicas razones para la acción que son relevantes para una persona son aquellas que ya tiene o aquellas a las que podría llegar por medio de una deliberación cabal a partir de sus actuales deseos, preferencias y evaluaciones (si dispusiera de plena información). Véase Bernard Williams, «Internal and External Reasons», reproducido en su *Moral Luck* (Cambridge: Cambridge Univ. Press, 1981), págs. 101-113. Esas razones internas se conectan con sus motivaciones actuales. Mas cuando una persona se pregunta a sí misma «¿Qué debería hacer?», no necesariamente se está preguntando qué es lo que mejor sirve a sus motivaciones presentes, o a las motivaciones que tendría si estuviera mejor informado. Puede que sepa que las motivaciones de los demás podrían diferir de y ser mejores

usando el balance neto de las razones que tiene para estimar el balance neto de las razones que hay, y luego, aplicar ese resultado a la estimación de la verdad. Pero la persona racional tratará de mantenerse alerta respecto de la posibilidad de que las razones de que dispone constituyan un indicador sesgado de la verdad y, consiguientemente, perfilará su estimación de la verdad según las razones que tenga. Si juzga que sus razones no son representativas y están sesgadas en una determinada dirección, procederá a una enmienda. Cuando la racionalidad evalúa las razones, se preocupa no sólo de la fuerza de las razones, sino también de su representatividad.

Éste es el sentido en el que puede decirse que la racionalidad entraña algún grado de autoconsciencia. No sólo las razones son evaluadas; también los procesos por medio de los cuales se adquiere, se almacena y se evoca información. Una persona racional tratará de estar alerta respecto de los sesgos de esos procesos, y se moverá para enmendar aquellos sesgos de que tenga noticia. Al evaluar la importancia de la información de que dispone, considerará también qué posible información *diferente* podría haber llegado a tener y cuán probable es que hubiera accedido a ella, dados varios hechos. Ésta es una de las lecciones del problema de los tres prisioneros[17]* y constituye un obstáculo adicional al proyecto de Rudolf Carnap de desarrollar una lógica inductiva para definir de un modo gene-

---

que las suyas en algunos respectos. Sus propias motivaciones pueden haber sido perturbadas por una infancia particularmente miserable o brutalizada, o quizá una particular situación haya decantado en ella ciertas motivaciones. Lo que ella quiere saber es cuáles son las *mejores* razones, y es posible que conceda algún peso a las opiniones de los demás al respecto. (¿Diría Williams que esto presupone un motivo interno preexistente, a saber; hacer lo que esté avalado por las mejores razones?)

* La gente que yerra en la solución del problema de los tres prisioneros —que infiere que cada uno de los dos prisioneros restantes tiene una probabilidad de un medio de ser ejecutado— lo hace como resultado de aplicar alguna regla o principio que les parece evidente, una regla insuficientemente sutil, o una regla que aplican sin la suficiente sutileza. Tenemos que darnos cuenta de que también nosotros podríamos hallarnos en esa situación —aun secundando una regla que fuera la mejor que podemos formular en el presente—, de modo que no fuéramos lo suficientemente perspicaces para usar una regla capaz de cancelar una razón afirmable (por ejemplo: «Usted sabía ya que uno de ellos no sería ejecutado, de modo que...»). Hay espacio abierto aquí para una investigación a la manera de Amos Tversky y Daniel Kahneman para determinar qué heurística general determinada usan quienes se dejan confundir por el problema de los tres prisioneros. Véase Tversky y Kahneman, «Judgment under Uncertainty: Heuristics and Biases», reproducido en *Judgment under Uncertainty: Heuristics and Biases,* comp. Daniel Kahneman, Paul Slovic y Amos Tversky (Cambridge: Cambridge Univ. Press, 1982), págs. 3-20. Se puede consultar también el resto de sus ensayos en este volumen.

ral el grado de confirmación de la hipótesis *h* dada la evidencia *e*.[18] Si las fuentes de información de una persona son tales que, aun si *p* fuera falsa, no podrían (o muy probablemente no podrían) proporcionarle esa información, entonces del hecho de que no reciba esa información no podrá concluir la persona que *p* es verdadero. Tiene que considerar que sus fuentes de información están sesgadas en favor de *p*.[19] Más adelante tendré ocasión de volver sobre el asunto de los sesgos de las razones.

Reglas de racionalidad

Tradicionalmente, los filósofos han tratado de formular reglas para la creencia racional, para la inferencia y la aceptación racionales de conclusiones deductivas y para modos racionales (no deductivos) de adquirir una creencia. Buscan reglas de apariencia atractiva, recomendables a la razón por su contenido y por su capacidad para generar las inferencias y las creencias en las que mayor confianza ponemos. Buscan reglas cuya aplicación permitiera afinar nuestros modos de acceder a y evaluar creencias.[20] No obstante, si la racionalidad de una creencia está en función de la efectividad del proceso que la produce y la mantiene, entonces no hay garantía alguna de que los procesos óptimos empleen reglas de apariencia atractiva. Al contrario, esos procesos podrían entrañar una competición entre reglas y procedimientos rivales, el peso de cada uno de los cuales vendría determinado (de acuerdo con específicos procedimientos de evaluación) por la historia pasada de la participación exitosa de cada regla en la realización de predicciones e inferencias. Ninguna de esas reglas o procedimientos necesita tener una apariencia razonable, sino que, constantemente modificadas por retroalimentación interactiva entre ellas, las reglas podrían actuar de consuno para producir resultados que satisfarían criterios externos deseables (como la verdad). La teoría de este proceso, así pues, no estaría constituida por un pequeño conjunto de reglas cuya aparente razonabilidad resultara detectable, de modo que una persona pudiera entonces aplicarlas de manera viable, sino por un programa computacional que simularía ese proceso.[21] Más radicalmente: es posible que no haya reglas simbólicamente representadas, ni siquiera en una competición ponderada, sino que cada «regla» surja como una regularidad de conducta de un sistema de procesamiento distribuido paralelamente, un sistema cuya matriz de los pesos que determinan la activación de vectores de salida o intermedios sea re-

petidamente modificada por alguna regla de corrección de errores.[22] Si los procesos más efectivos para alcanzar objetivos cognitivos son de esta clase, entonces el tipo de reglas normativas que han tratado de formular los filósofos para distinguir la racionalidad de la creencia está condenada al fracaso. Tales reglas ni siquiera serían partes del proceso, y la consciente aplicación de las mismas no sería el mejor camino hacia la creencia verdadera (o estimable por otros motivos).

Los filósofos se han vuelto tecnológicamente obsoletos en el estudio de los procesos fiables de adquisición de creencias. Serán reemplazados por científicos cognitivos y computacionales, por especialistas en inteligencia artificial y otros.* Nuestra comprensión progresará, pues, pero la naturaleza de esa comprensión cambiará: las simulaciones computacionales vendrán a reemplazar a teorías que presentaban reglas estructuralmente relevantes, reglas investidas de una apariencia de validez y que la gente podía apreciar y aplicar.[23] Esto podría sernos útil —produciremos máquinas para ejecutar tareas intrincadas—, pero no será lo que los filósofos esperaban; reglas y procedimientos que podemos aplicar por nosotros mismos para mejorar nuestras propias creencias, reglas y procedimientos *examinables* —tomo el término de Wittgenstein— que podamos adoptar y entender globalmente y que nos proporcionen una descripción estructuralmente reveladora de la naturaleza de la racionalidad. (Considérese la diferencia entre un consejo estratégico tradicional para jugar al ajedrez y un programa que aprendiera *solamente* a través de los resultados de sus movimientos pasados merced a algún procedimiento retroalimentador que registrara los éxitos. Una máquina así programada podría acabar jugando muy bien, pero nosotros no entenderíamos por qué jugó como jugó en cada partida; de su juego no podríamos aprender ninguna regla particular secundable para mejorar nuestro propio juego. El elaborado sistema de pesos al que habría llegado el programa quizá no sea formulable con

* La bibliografía que da la pauta de esta transición es enorme. Para reflexiones generales, véase Clark Glymour, «Artificial Intelligence Is Philosophy», en *Aspects of Artificial Intelligence*, comp. James Fetzer (Dordrecht: Kluwer, 1988), págs. 195-207. No me estoy refiriendo aquí a un asunto de delimitación de las disciplinas académicas. Evidentemente, alguna gente educada filosóficamente desplazará su trabajo hacia esas áreas, y personas que en el futuro se habrían dedicado a la filosofía se entrarán en esas áreas. Lo que resulta interesante aquí es el cambio de naturaleza del objetivo teórico y del tipo de comprensión que de ello resulta. Con todo, los análisis conceptuales de los filósofos acerca de lo que anda en juego en varios ámbitos y en diversas tareas podrían ayudar a los especialistas en inteligencia artificial y en ciencia cognitiva a evitar pautas de investigación condenadas al fracaso.

conceptos que nos resulten accesibles de una manera inteligible para nosotros.) Las máquinas en las que moran esos programas, no obstante, podrían convertirse en una útil ayuda externa para mejorar nuestras creencias. (Yo puedo creer la respuesta que da una calculadora sin entender por qué es la respuesta correcta.)

Supongamos que los procesos más fiables para llegar a una creencia implican esos procesos registradores del éxito pasado, así como la continua revisión de los pesos; las únicas reglas son las reglas que determinan qué entra en la competición, con qué fuerzas y cómo han de ir modificándose esas fuerzas según los resultados averiguados por el uso del vencedor de la competición.* (He aquí una regla: si ustedes han errado por lo alto en su blanco, modifiquen ligeramente hacia abajo todos los pesos positivos que interactúan con cantidades positivas; si ustedes han errado por lo bajo, modifiquen ligeramente hacia arriba todos los pesos positivos que interactúan con cantidades positivas; sigan con esas modificaciones hasta que den exactamente en el blanco, luego paren. Sin duda se trata de una forma admirable de regla —instrucciones específicas les dirán exactamente qué modificaciones hay que ir haciendo en los pesos—, pero no es un principio de creencia racional del tipo de los que antaño se buscaban. Ni lo es tampoco el «algoritmo de la brigada del cubo» para el suministro de crédito,[24] aunque trata con entidades más parecidas a las reglas que la regla delta.) Aun así, podría decirse, los principios de los filósofos tienen una función iluminadora, a saber: describir el *output*, el producto, de esos procesos de retroalimentación (procesos, registradores de éxitos pasados, que modifican sus pesos de acuerdo con esos éxitos en las predicciones y en las inferencias presentes). Algunos principios pueden *definir* objetivos cognitivos, y por lo mismo, definir el blanco al que apuntan los procesos, mas los principios (existentes) no podrían describir ese producto más iluminadoramente. No podemos saber de antemano si varios principios filosóficos (que no se limiten a definir los objetivos cognitivos) describirán adecuadamente el producto de los procesos más efectivos para conseguir esos objetivos.

Consideremos, por ejemplo, el requisito normativo, frecuentemente propuesto, de que el cuerpo de creencias de una persona sea co-

* «Mas, ¿confiaríamos en el veredicto de tal mecanismo si entrara en conflicto con nuestra intuición acerca de un caso o una proposición determinada?» Yo, desde luego, a la hora de elegir mi mejor movimiento, y aun sin ser capaz de entender su sentido, confiaría en los consejos de una máquina de jugar al ajedrez muy experimentada cuyos movimientos fueran el resultado de ponderaciones modificadas de acuerdo con alguna regla de corrección de errores.

herente y esté clausurado deductivamente (que cualquier consecuencia lógica de lo creído sea también creída). Quizá los procedimientos más efectivos para conseguir una elevada proporción de verdades (y relativamente pocas falsedades) arroje un conjunto de creencias que es incoherente. Así que, si hay que conservar esa elevada proporción de verdades, es mejor que el conjunto de creencias no esté clausurado deductivamente. Exigir de antemano que no se usen reglas para generar creencias de las que se sabe que (en determinadas circunstancias) llevan a incoherencias podría impedirnos llegar a adquirir un número muy grande de creencias verdaderas. Cuando, en cambio, se usan tales reglas hay que tomar medidas para limitar las incoherencias que pudieran aparecer. Podemos buscar una manera aceptable de evitar la incoherencia, pero entretanto nos embarcaremos en el control de daños aislando incoherencias y tomando medidas para no inferir todos y cada uno de los enunciados arbitrarios que se siguen de las contradicciones explícitas.[25] En la vida cotidiana lo hacemos con ecuanimidad —reconocemos paladinamente nuestra fiabilidad y afirmamos que, sin duda, una de nuestras creencias es falsa—. La ciencia incorpora un fuerte impulso para conseguir la coherencia, pero también los científicos se andan con reservas y rechazan la obligación de renunciar a muchas predicciones exactas generadas por una teoría reconocida por su capacidad para generar también incoherencias o valores imposibles (como lo atestigua el uso de la «renormalización» en los cálculos de la mecánica cuántica).

Sin embargo, a la empresa filosófica tradicional que trataba de formular principios explícitos para la creencia racional aún le queda un papel por desempeñar —un papel más modesto—. El veredicto de un determinado principio explícito *P*, según el cual es racional (o irracional) creer un enunciado *q*, puede ser uno de los productos de un proceso de creencia. El veredicto del principio *P* no será por sí mismo determinante de la creencia, pero entrará en ulteriores estadios del proceso de formación de la creencia y tendrá su peso (en la formación de la creencia futura) modificado por el resultado percibido de aceptar o rechazar ese enunciado *q* y actuar consiguientemente. Tal regla o principio *P* podría ser una unidad de procesamiento, una entre muchas otras dentro de un sistema de procesamiento paralelamente distribuido, y su producto —sea un veredicto de racionalidad o de irracionalidad, una probabilidad bayesiana, o lo que fuere— podría verse propagado y acabar desempeñando un papel en la activación de alguna unidad contigua hasta que, finalmente, se lograra un resultado, creencia o no, y se averiguara su resultado

ulterior.* El peso asignado a las conexiones del principio *P* con componentes adicionales pertenecientes a este sistema global sería entonces modificado por la retroalimentación procedente de los resultados últimos de acuerdo con alguna regla de aprendizaje. (Quizá llegarían a verse modificados algunos detalles del principio *P*, produciéndose así un nuevo principio *P'*.)

Las observaciones precedentes no significan una aceptación por mi parte de los programas (estrictos) de investigación de los propugnadores del procesamiento paralelamente distribuido, ni están ligadas a ese programa a pesar de la popularidad de que gozan actualmente esas teorías conexionistas.[26] Más adelante supondré que reglas específicas libran un producto dentro de un complicado proceso de retroalimentación, y no me preocupa particularmente aquí el que esas reglas surjan como regularidad de un sistema de procesamiento paralelamente distribuido o estén ellas mismas representadas simbólica y explícitamente. Para mis propósitos, lo más sugerente del procesamiento paralelamente distribuido es el marco general que presenta: múltiples unidades que se progrealimentan en unidades ulteriores (cuyas acciones están determinadas por el alimento que reciben) de acuerdo con una matriz de pesos que se

* Los propugnadores de los sistemas de procesamiento paralelamente distribuido han tendido a sostener que las reglas surgen como resultado de la sucesiva asignación de pesos y no necesitan ser representadas simbólicamente por doquier en tanto que tales —surgen, me siento tentado a decir, por un proceso de mano invisible—. Las «reglas» se incorporan aquí a una pauta de conectividad entre unidades procesadoras. Otros lugares en los que se pueden buscar reglas son: la función del producto por cada unidad, la regla de propagación para pautas propagadoras de actividad a través de la red, la regla de activación para combinar los *inputs* que chocan en una unidad con el estado presente de esa unidad, combinación que genera un nuevo nivel de activación para la unidad, y una regla de aprendizaje o de corrección de errores merced a la cual se van modificando con la experiencia las pautas de conectividad. (Sigo aquí la lista de aspectos de un modelo de procesamiento paralelamente distribuido que se ofrece en D.E. Rumelhart, G.E. Hinton y J.L. McClelland, «A General Framework for Parallel Distributed Processing», en *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, comp. Rumelhart y McClelland [Cambridge, Mass.: M.I.T. Press, 1986] [trad. cast.: *Introducción al procesamiento distribuido en paralelo*, Madrid, Alianza, 1992], pág. 46.) En este último lugar, en la regla de aprendizaje, es donde puede esperar encontrarse alguna representación simbólica. En cualquier caso, mis observaciones sólo pretenden mostrar cómo, dentro de un marco general de algún modelo de procesamiento paralelamente distribuido, una regla podría llegar a cumplir una función; no parto del estricto supuesto de que hay representación alguna de reglas, sólo pautas de conectividad. Para varias críticas de la adecuación de las teorías psicológicas conexionistas que no representan simbólicamente las reglas, véanse los ensayos en *Connections and Symbols*, comp. Steven Pinker y Jacques Mehler (Cambridge, Mass.: M.I.T. Press, 1988).

modifica mediante alguna regla retroalimentativa (de corrección de errores). Algunas de las unidades dentro de esa red pueden ser *reglas*. Pero exhibirán relaciones entre ellas parecidas a las relaciones que se dan entre unidades que no son reglas dentro de una estructura de procesamiento paralelamente distribuido.

En ese marco general, podemos decir que el filósofo tradicional esperaba formular un principio cuyo *output* o producto determinara completamente el estadio ulterior de la creencia al ser el único *input* de algún peso para la unidad de creencia. En un sistema de procesamiento paralelamente distribuido, un componente por el estilo de un principio podría llegar a desempeñar un papel tan destacado sin necesidad de proceder de esa manera; con el tiempo, los pesos de todos los demás componentes (independientes) cuyos productos chocaran con la unidad de creencia podrían llegar a ser cero. (De modo que este marco general deja margen para la esperanza del filósofo tradicional, pero no depende de ella.) Y aun sin llegar a una regla tan fuerte, un principio explícito puede sernos útil, y su veredicto, parte del proceso (más) fiable para adquirir una creencia. Los principios explícitos deberían tomarse *cum grano salis*.

Las varias máximas metodológicas que presentan los libros de filosofía de la ciencia podrían incrustarse en ese marco general. Algunas de esas máximas podrían ser componentes del sistema, y tendrían un peso gradualmente modificado; algunas máximas podrían surgir como descripciones de la operación del sistema, más que ser componentes de él.[27] Me refiero a máximas como las siguientes: el apoyo evidencial es mayor cuanto más variada es la evidencia (en donde la variedad se juzga de acuerdo con categorizaciones incorporadas en otras creencias); son preferibles las hipótesis más simples; una teoría o una hipótesis hereda el apoyo evidencial de la teoría que viene a substituir; las predicciones de nuevos fenómenos añaden más credibilidad que las derivaciones de fenómenos conocidos; hay que evitar las hipótesis *ad hoc;* hay que controlar todas las demás variables que podrían afectar al resultado; cuanto más precisa la predicción exitosa, más confirmada resulta la hipótesis.

Ya hemos visto cómo un principio explícito puede desempeñar un papel útil en la producción de una creencia verdadera (o de una creencia que satisface algún otro objetivo cognitivo apropiado). Pero ¿puede un principio explícito ayudarnos a *entender* la naturaleza de la creencia racional? Una manera de entender la racionalidad de una creencia consiste simplemente en contemplarla como el resultado de cierto tipo de proceso —por ejemplo, de un proceso paralelamente distribuido (por poner el peor caso que pueda presentarse

para la intelección proporcionada por principios explícitos)—. O podemos entenderla, con más detalle, como el resultado de este particular proceso, definidos todos los parámetros, funciones y reglas de transición. (A modo de comparación: la actual vida animal y vegetal es lo que surgió de las continuas operaciones de estos particulares procesos darwinianos; la distribución actual de la salud y de la renta en una sociedad de mercado es lo que surgió de la continua operación de estos particulares procesos.) Con todo, podemos querer entender algo más, a saber: las pautas normalmente mostradas por los resultados de esos procesos, en la medida en que exhiban una pauta describible. ¿Cómo funcionan los organismos normales? ¿Cuál es la pauta de distribución de la renta y de la salud y con qué factores se correlaciona? ¿Qué aspecto tienen las creencias racionales? Ahora bien; si la racionalidad apunta a la verdad, los principios para llegar a la creencia racional quizá no sean la mejor descripción del aspecto de las creencias racionales. La mejor descripción breve del aspecto del conjunto de creencias racionales podría ser nuestra más compendiada descripción del conjunto de verdades —la más amplia que podemos hacer simplemente, sin enfrentarnos a problemas de diagonalización—. Éste no coincidirá con el conjunto de creencias racionales —a menudo creemos falsedades, y hay algunas verdades en las que no creemos—, pero podría ser la mejor descripción breve e inteligible que podemos ofrecer, y una parte de la falta de encaje quedará reducida por el hecho de que somos *nosotros* quienes estamos ofreciendo una descripción, de manera que nuestra descripción del conjunto de verdades incluirá también errores en la dirección de nuestro conjunto de creencias.

Lo que acaso deseemos, empero, no es ninguna de estas cosas —ni una descripción de las pautas que se adivinan en nuestras creencias racionales presentes, ni una descripción del proceso del que surgieron las creencias actuales—. Quizá deseemos más bien una descripción de la relación de nuestras creencias racionales actuales con su fuente, una descripción estructuralmente reveladora y relativamente breve del modo en que el contenido y la estructura de nuestras creencias racionales actuales se relacionan con el contenido y la estructura de aquello de lo que proceden. ¿Cuál es la pauta de *esa* conexión? Quizá no haya nada que entender aquí. Es posible que, suburdida y establecida por el proceso competitivo de retroalimentación según los éxitos, esa conexión no muestre pautas reveladoras, ni siquiera una cruda aproximación de las mismas. (Entonces no habrá nada que no entendamos; simplemente no habrá nada que entender.) Pero si la conexión muestra una pauta, o estadios sucesi-

vos de la misma, entonces deberíamos ser capaces de formular esa pauta con algún principio explícito. Tal principio puede también ser de ayuda a la hora de decidir qué credibilidad dar a las cosas que nos cuentan otros mostrando la relación entre lo que dicen y aquello en que se basa lo que dicen.[28]*

Me gustaría describir ahora *un* componente de la red de muchos componentes con varios pesos (de acuerdo con alguna regla de retroalimentación) que habrá de determinar nuestra creencia. Será útil empezar con el teorema de Bayes. Este teorema, muy fácilmente derivable de los axiomas de la teoría de la probabilidad, dice de qué debería depender la probabilidad de una hipótesis *h* dada la evidencia *e*. De una fórmula compleja podemos convertir este teorema en un enunciado intuitivo.

¿Cuán probable es la hipótesis *h* con los datos *e*? Bien, ¿cuán probable es que los datos aparezcan como resultado de la (verdad de la) hipótesis? ¿Cuán probable es que el hecho de que *e* sea el caso se deba a la vigencia de *h*? No se trata simplemente de la probabilidad de que, siendo *h* verdadera, aparezca *e*. Llamemos por el mo-

* Hay otras posibles vías para evaluar la fiabilidad de otro, vías que no se refieren a las conexiones pautadas que delimitan sus creencias; por ejemplo: estadísticas sobre su fiabilidad. Pero, en general, serán vías de tránsito difícil. Sin embargo, la pregunta «¿Por qué piensa usted así?» podría admitir una respuesta relativamente fácil y reveladora. El que nos dijera «He tenido mucha experiencia con estas cosas, y puedo hablar» también podría ser de ayuda. En términos corrientes, nos dice que sus pesos actuales, y por lo tanto sus respuestas, han sido modeladas por y se benefician de una gran cantidad de retroalimentación. Esto constituiría una garantía, siempre que pensáramos que el enunciado es de tal tipo que un sistema así podría aprender a reconocer su verdad.

Un proceso puede ser fiable en punto a acabar llegando a la verdad sobre un asunto —acabar en una creencia verdadera en un alto porcentaje de casos—, aunque esto consume mucho tiempo. Si a lo largo del camino que se recorre en este proceso se llega a creencias, éstas se modifican retroalimentativamente, etc., entonces muchas de estas primeras creencias habrán sido falsas. Una creencia presente puede haber surgido merced a un proceso fiable —un proceso que acaba en una verdad un elevado porcentaje de veces—, pero en una etapa del proceso que no le concede una probabilidad de ser verdadera mayor que 1/2. ¿Es racional creer ahora en *ella*? Una sugerencia es apoyar el proceso en algún mecanismo externo y abstenerse de creer algo hasta que el mecanismo descanse establemente en la creencia. (Nuestro procedimiento fiable entonces es *este* procedimiento.) Mas esa opción podría no estar a nuestra disposición. Quizá el proceso puede funcionar sólo en nuestras cabezas, y la retroalimentación funcionar sólo respecto de aquello que creemos realmente. Poner en marcha ese proceso, entonces, un proceso que acabe siendo fiable, requerirá que tengamos realmente muchas creencias por el camino que no son fiablemente verdaderas. (La ciencia sería un ejemplo de esto si los científicos tienen que creer en sus resultados presentes para poder hacer ciencia normal, generar anomalías que exijan una nueva teoría, etc.)

mento a *esta* probabilidad prob(*e*:*h*). Es decir, cuán probables son los datos *e dado* que *h* es verdadera —hablo aquí laxamente—, cuán probable sería *e si h fuera* verdadera.

Para hallar la probabilidad de que los datos *e surjan* por la vía de la hipótesis *h*, no sólo tenemos que considerar la prob(*e*:*h*), sino también la probabilidad de *h*. Que *e* surja de *h* puede ser improbable, por alta que sea la prob(*e*:*h*), si la prob(*h*) es muy baja. La probabilidad de que suene una trompa fuera en este momento sería muy alta si hubiera una cábala de Centuriones Alfa escondidos investidos de grandes poderes y que lo que más desearan fuera que yo oyera sonar una trompa precisamente ahora. La probabilidad de esta última hipótesis es, sin embargo, minúscula, y la hipótesis no gana significativamente en probabilidad por su capacidad para dar cuenta de este particular estruendo de trompa. La prob(*e*:*h*) x prob(*h*) representa la probabilidad de que *e surja de la verdad de h*, la *probabilidad absoluta* de que *e* surja de la verdad de *h*.

Pero ¿qué nos dice esto de la probabilidad de *h*, o de la probabilidad que tendría *h* dado *e*? Conocemos la probabilidad absoluta de que *e surja vía h*, pero también hay una oportunidad de que *e* surja de algunas *otras* hipótesis *h*2, *h*3, *h*4, ... Queremos atender no sólo a la probabilidad absoluta de que *e* surja de *h*, sino también a la probabilidad *relativa* de que lo haga. ¿Qué *porcentaje* de las probabilidades de todas las (otras) vías por las que *e* hubiera podido surgir representa la probabilidad de que surja vía *h*? (Matemáticamente, es mejor suprimir ese «otras» entre paréntesis.) La probabilidad *relativa* de que *e* surja de *h* es el cociente entre la probabilidad absoluta de que *e* surja vía *h* y la suma de las probabilidades absolutas de que *e* surja en *todas* las vías por las que podría surgir. Lo que parece decirnos cuán probable es *h* dados los datos *e* es la probabilidad relativa de que *e* surja de *h*. Rebauticemos la hipótesis *h* de la que nos hemos estado ocupando y llamémosla *h*1. Entonces la probabilidad de *h*1 *fundada en e* parece ser:

$$\frac{\text{prob}(e{:}h1) \times \text{prob}(h1)}{\sum_{(i=1)}^{n} \text{prob}(e{:}hi) \times \text{prob}(hi)}$$

Y esta fórmula parece decirnos cuán probable es *h*1 dado *e*. Pues nos dice cuán probable es que *e* surja de *h*1, en comparación con todas las vías por las que *e* hubiera podido surgir. De manera que parece que tenemos una derivación intuitiva, o al menos una explicación, del teorema de Bayes.

Pero a lo que hemos llegado no es exactamente al teorema de Bayes. Nuestras probabilidades de que si una hipótesis fuera verdadera daría lugar a la evidencia *e*, prob(*e*:*h*), no son las probabilidades condicionales del teorema de Bayes. Estas últimas son simplemente las probabilidades de que la evidencia ocurra dado que se cumpla la hipótesis, *dé o no lugar* la hipótesis a esta evidencia, haya o no haya alguna conexión (probabilística) subjuntiva entre la hipótesis y la evidencia. A lo que hemos llegado es a una versión causalizada o subjuntivizada del teorema de Bayes. Para poner de relieve que esas probabilidades de los subjuntivos no son las probabilidades condicionales estándar, simbolicemos la probabilidad de que si *h* fuera verdadera, *e* sería verdadera, así: prob($h \rightarrow e$).

La versión sobredicha necesita cierta modificación, no sólo tipográfica. Pues aunque hemos considerado las diferentes hipótesis que hubieran podido dar lugar a la evidencia *e*, no hemos considerado aún la posibilidad de que *e* hubiera podido ocurrir espontáneamente, sin causa (probabilística) o sin hecho hipotético generador. También hay que tomar en cuenta esa posibilidad. (Podemos considerar esta hipótesis azarosa acerca de *e*, que simbolizamos como *Ce*, simplemente como la negación de que haya alguna *hi* tal que *hi* genere probabilísticamente *e*.) Así, nuestra fórmula se convierte en:

$$\text{medida } (h1/e) = \frac{\text{prob}(h1 \rightarrow e) \times \text{prob}(h1)}{\sum_{(i=1)}^{n} \text{prob}(hi \rightarrow e) \times \text{prob}(hi) + \text{prob}(Ce)}$$

¿Qué mide esta medida?[29] ¿Fija esta fórmula la medida de *h*1 dada *e* como una probabilidad condicional, interpretada esa probabilidad como apostar con ventaja en una apuesta condicional, o fija más bien la fórmula alguna otra cantidad, como por ejemplo el grado de apoyo que ofrece *e* a *h*1, o la probabilidad de que si *e* fuera verdadera, *h*1 sería verdadera, prob($e \rightarrow h1$)? Una cuestión preliminar a investigar es cuáles son las propiedades del cociente en el lado derecho de la fórmula, si el cociente se comporta como una probabilidad, etc.

Hay una abundante literatura filosófica dedicada a las inferencias explicativas, las «inferencias hacia la mejor explicación».[30] Decir que hay un tal principio de *inferencia* es sostener que la bondad de *h* como hipótesis explicativa basta para determinar la credibilidad de *h* como creencia. Lo cierto, sin embargo, parece algo más dé-

bil, a saber: que la bondad de *h* como hipótesis explicativa es uno de los factores que entra en la fijación y en la evaluación de la credibilidad de *h*.[31] El cociente arriba enunciado parecería entonces útil a la hora de definir y analizar ese factor. La bondad de *h* como explicación de *e* dependerá, al menos, de la probabilidad de que si *h* fuera verdadera lo fuera también *e*, y de la probabilidad previa de *h*; resulta plausible que la dependencia respecto de esas probabilidades lo sea respecto de su producto. *Ceteris paribus,* la hipótesis *h* suministra una mejor explicación que otra hipótesis *hi* si el producto para *h* es mayor que el producto para *hi*.

No siempre podríamos querer inferir lo que sea la mejor explicación, sin embargo. Supongamos que hubiera ocho posibles *hi*, donde la prob($hi \rightarrow e$) fuera la misma para cada hipótesis *hi*, y prob(*h*1) fuera sólo una pizca mayor que un octavo, mientras que las probabilidades de *h*2, ..., *h*8 fueran cada una de ellas una pizca (menor) más pequeñas que un octavo. Es posible que *h*1 sea la mejor explicación, pero sin embargo no es una explicación tan buena. Eso es lo que indica el pequeño cociente de *h*1 tal como se mide en el lado derecho de nuestra ecuación. Evidentemente, los éxitos explicativos previos de una hipótesis serán importantes, y afectarán a su probabilidad previa, que entra en esta fórmula. También otros factores podrían influir en esa probabilidad previa. Así, es posible que el cociente en cuestión mida el grado de fuerza explicativa que confiere *e* a *h*1.* Al valorar el estatus explicativo de una hipótesis es pertinente considerar algo más que el hecho singular *e* que se trata de explicar: ¿qué hipótesis recibe el mayor apoyo explicativo de los hechos que han de ser explicados?[32] Dentro de una red con muchos componentes que se progrealimenta con varios pesos (como lo es una red de procesamiento paralelamente distribuido), este valor explicativo de *h* sería *uno* de los factores que se inyecta en la credibilidad global de *h*.

Imaginemos una red que incorpore una ponderación de muchos factores —incluidos las probabilidades bayesianas, el valor explicativo (representado por la fórmula bayesiana causalizada), las máximas de la metodología popperiana y una evaluación de los reveses

* Esto es, el grado de apoyo explicativo que *e* confiere a *h*1, dadas las hipótesis que han sido formuladas y de las que se tiene noticia. Alguna *otra* hipótesis $hn + 1$ que no habría aún sido formulada podría explicar también *e*, y esa hipótesis no está incluida en ningún factor del denominador de la fórmula, ni siquiera en el último factor (según el cual *e* ocurre por azar). De modo que el cociente es una medida del grado de apoyo explicativo que recibe *h*1, de *e*, *relativo a* nuestra presente formulación de hipótesis explicativas alternativas admisibles.

experimentados— y proceda a una progrealimentación que resulte en un *valor de credibilidad* para el enunciado *h*. Podemos entender esto como una evaluación ideal que sopesa debidamente todas las razones en favor y en contra de *h*.[33] Mi presentimiento es que, dentro de ese sistema, las hipótesis ricas se vuelven más ricas. Los datos nuevos no contarán como evidencia en favor de cualquier hipótesis que case con ellos. Serán acumulados por la más creíble de esas hipótesis, y así, engrosarán *su* credibilidad, no la de las demás hipótesis. (Esto podría ejemplificarse con un sistema en el que las hipótesis existentes participaran en una subasta por el apoyo de nuevos datos; las hipótesis que ya fueran ricas en credibilidad tendrían una ventaja en esa subasta.) Cuando la hipótesis más creíble hasta ahora resulta rechazada por alguna razón, los datos que le daban apoyo se ponen a disposición de otra hipótesis.[34]

Este sistema de procesamiento es un sistema de aprendizaje; sus pesos son modificados por retroalimentación. ¿De dónde viene esa retroalimentación? Las correcciones de las predicciones y de las expectativas del sistema pueden proceder de un maestro externo que da entrada a valores corregidos, o de un *input* sensorial distinto del predicho, o del registro de desarmonías internas que reducen la fuerza de cada uno de los componentes.[35] La regla concreta de corrección de errores variará de un sistema a otro, quizá incluso, dentro de un mismo sistema, de tarea a tarea, y acaso la regla misma compita con otras reglas de corrección de errores en un sistema de retroalimentación sujeto a alguna regla ulterior de corrección de errores, o incluso a una de las reglas que compite con ella.

Considerada una persona como un sistema de este tipo, presenta varios aspectos relevantes para evaluar su racionalidad global. ¿Tiene un buen conjunto de pesos y fuerzas? ¿Posee un buen sistema de detección que registra la información adecuada para los propósitos de la retroalimentación? ¿Posee una regla de corrección de errores que modifica los pesos de un modo eficiente? Y (quizá) ¿tiene algún procedimiento para revisar la estructura global de la red? El conjunto del sistema puede ser racional sin que ninguno de esos aspectos sea óptimo. Quizá haya reglas de corrección de errores que converjan más rápidamente en pesos adecuados que no necesiten corrección, pero el sistema será (de algún modo) racional si la regla existente hace al menos (pequeñas) modificaciones en la dirección correcta y con el signo correcto. Lo que hay que considerar es el modo en que los componentes se relacionan entre sí en el funcionamiento del sistema como un todo.

Hasta ahora hemos imaginado un sistema que genera valores de

credibilidad, dando una puntuación para cada enunciado *h* objeto de evaluación. ¿Cómo hay que usar el resultante valor de credibilidad de *h* para llegar a una creencia sobre *h*? ¿Cuál será la regla de aceptación de *h*?

La primera regla de aceptación es ésta:

> **Regla 1:** No creas *h* si algún enunciado alternativo incompatible con *h* tiene un valor de credibilidad mayor que el de *h*.

Los valores de credibilidad se diferencian de las probabilidades en que el valor de credibilidad de *h* y el de no-*h* no requieren ser sumados y resultar en un valor fijo. No obstante, cuando no-*h* tiene un valor de credibilidad mayor que el de *h*, la regla 1 nos insta a no creer *h*. (Pero, puesto que en la determinación del valor de credibilidad entran varios factores, como por ejemplo la fuerza explicativa, no es cuestión de suponer que la relativamente indefinida no-*h* tendrá siempre un valor de credibilidad mayor que el de *h*.)

Obsérvese que esta regla se aplica también cuando el enunciado incompatible con *h* no es alguna hipótesis alternativa situada a su nivel, sino más bien alguna contraevidencia *g* de *h*, estrictamente incompatible con *h* —un falsador popperiano—. En ese caso, si el valor de credibilidad de *g* es mayor que el de *h*, entonces *h* no será aceptada. Con todo, es posible que esta aparente disconfirmación no sea concluyente, pues el valor de credibilidad de *g* puede ser más pequeño que el de *h*, y podría ocurrir que se viera disminuido por el hecho de que *h* aumentara el peso de algún socavador de *g*.[36]

La regla 1 elimina algunos enunciados y nos deja con un conjunto de candidatos para ser creencias actuales, a saber: aquellos cuyo valor de credibilidad no esté rebasado por el de algún enunciado incompatible con ellos. Entre esos candidatos admisibles, ¿cuál será creído? ¿Los creeremos todos, o creeremos más bien todos aquellos que subsistan después de aplicar algún procedimiento de poda adicional para eliminar incoherencias? No creo que debamos seguir esta política maximalista sobre la creencia. En este punto, creo antes bien que deberíamos seguir un cálculo de teoría de la decisión sobre la deseabilidad de sostener tal creencia, un cálculo que incluyera objetivos y utilidades *prácticos,* no simplemente cognitivos. (No quiero con ello decir que debiéramos hacer explícitamente ese cálculo; sino que deberíamos actuar de acuerdo con él, revisando nuestra acción si nos apercibimos de una desviación de ese cálculo.) Ésta es nuestra segunda regla de aceptación.

Regla 2: Cree (una admisible) *h* sólo si la utilidad esperada de creer *h* no es menor que la utilidad esperada de no tener ninguna creencia sobre *h*.*

Tenemos, así pues, un procedimiento en dos etapas: el primero escarda los valores de credibilidad más bajos, y el segundo, de entre los enunciados que subsisten, determina la creencia considerando las consecuencias (latamente concebidas) de tal creencia. Hay muchas cosas que hablan en favor de tal procedimiento. Secundándolo, queda excluido que una persona mantenga creencias cognitivamente inferiores (según el valor de credibilidad). Sin embargo, no está obligada a mantener una creencia sobre este asunto. El que la mantenga o no quedará determinado por sus objetivos y por sus fines prácticos, teóricos y sociales.

Consideremos de nuevo el caso de la madre del delincuente convicto. Supongamos que el valor de credibilidad que ella concede a la culpabilidad de su hijo es mayor que el que concede a su inocencia —el mayor conocimiento de su hijo del que ella dispone no consigue invertir los valores de credibilidad hallados por otros—. Creer en su culpabilidad, sin embargo, aun si esa creencia fuera solvente, le reportaría a ella una gran desutilidad. De acuerdo con el principio en dos etapas propuesto aquí, es irracional para ella creer en la inocencia de su hijo. (Que su hijo sea inocente no es algo que ella pueda creer racionalmente —la regla 1 excluye tal creencia—.) Sin embargo, no es irracional para ella *no* creer que su hijo es culpable; no es irracional para ella no sentar ninguna creencia sobre el asunto. (¿Qué ocurre, empero, si la vida de la madre se degradara miserablemente en caso de no creer en la inocencia de su hijo? Podemos invocar aquí nuestra anterior distinción. Que el hijo sea inocente no es algo que ella pueda creer racionalmente; pero creerlo podría ser lo mejor, y por lo tanto, lo más racional para ella.) ¿Qué si la evidencia de culpabilidad fuera *abrumadora*? Mas ¿cuán estrictos son los criterios que deben ser satisfechos antes de que podamos creer que

* Una regla más estricta requeriría que la utilidad esperada de creer *h* fuera mayor que la de no tener creencia alguna sobre *h*. La diferencia estriba en si deberíamos creer *h* cuando hay un empate exacto entre la utilidad esperada de creer *h* y la utilidad esperada de no tener creencia alguna sobre *h*. Pudiera pensarse que en este último caso deberíamos creer porque creer la verdad tiene un valor —mas ¿no estaría ya ese valor incluido en el cálculo de utilidades que, por definición, arroja un empate?—. O pudiera pensarse que en tal caso no deberíamos creer porque tener una creencia conlleva costes (tenemos una capacidad limitada de almacenamiento, una limitada energía mental, etc.) —mas ¿no habrían sido ya incluidos esos costes en el cálculo de utilidades?—.

una determinada cosa puede, en parte, convenirnos? Nosotros decidimos qué nivel de evidencia, qué cantidad de credibilidad, hace falta para convencernos; y la altura a la que situemos ese nivel puede depender de muchos factores, incluida la utilidad que nos reporta a nosotros y a la sociedad el tener ciertas creencias. (No es necesario que la madre fije el criterio de manera tal, que nada pueda satisfacerlo. Quizá pueda decidir creer en la culpabilidad de su hijo sólo si tiene la experiencia directa y sobrenatural de un mensaje de Dios comunicándole la culpabilidad de su hijo.) Que podamos elegir el criterio no significa que podamos elegirlo arbitrariamente. Es posible que anden implicados aquí los principios y las exigencias de consistencia: los casos que la persona contempla como similares podrían tener que satisfacer el mismo criterio; el rigor del criterio podría tener que variar directamente (y nunca inversamente) con las posiciones de los casos a lo largo de una dimensión relevante (por ejemplo, la cantidad de utilidad que va con la creencia), etc.*

Obsérvese que un enunciado queda excluido como candidato a la creencia en la primera etapa sólo si su valor de credibilidad es menor que el de algún enunciado incompatible.[37] Con todo, no es necesario que un candidato a la creencia tenga un valor de credibilidad global máximo. Su valor de credibilidad puede ser menor que el de algún otro enunciado no incompatible. Obsérvese asimismo que cuando se rechaza para la creencia un enunciado en la segunda etapa —creer no es deseable por razones prácticas—, tal enunciado sigue activo en la primera etapa y puede excluir enunciados de menor credibilidad incompatibles con él.** Bajo este principio, la creen-

* El público del juzgado puede decir a la madre: «Nosotros consideramos que la evidencia es suficiente como para que sea *irracional* no albergar la creencia de que su hijo es culpable». La madre y el público coinciden en que el enunciado de menor credibilidad (que su hijo es inocente) no puede ser creído, pero discrepan acerca del valor de credibilidad necesario para la creencia en su culpabilidad. La madre puede preguntar cómo decidieron dónde fijar el umbral de suficiencia evidencial. Si, como parece plausible, la fijación del umbral por parte del público estaba determinada por los efectos promedio resultantes de albergar ciertas creencias acordes con los dictados de ese umbral, ¿no podría acaso la madre replicar que los efectos para ella no son los efectos promedio, y que si los efectos promedio *del público* tuvieran la magnitud de los de ella, también el público habría fijado un umbral más estricto? ¿Por qué habría ella de albergar una creencia que obedeciera al criterio apropiado para la diferente situación del público? (Después de todo, los miembros del jurado fijan apropiadamente su umbral en un lugar distinto del que fija el asistente promedio a las sesiones del juzgado.)

** Supongamos que p tiene un valor de credibilidad mayor que q, y q un valor de credibilidad mayor que r; supongamos también que p es incompatible con q, y q con r, pero que p no es incompatible con r. ¿Es r un candidato inadmisible para

cia es una combinación de lo teórico y de lo práctico, gozando lo práctico de prioridad (lexicográfica).

Que la primera etapa sea una etapa eliminativa captura el presentimiento de que, en lo atinente a la racionalidad de la creencia, la «irracionalidad» es la noción primaria —en expresión (¿sexista?) de J.L. Austin, la que lleva los pantalones—. Resulta claro que hay muchas cosas que es irracional creer. Es menos claro que algunas creencias sean tan creíbles que vengan mandadas por la racionalidad, de modo que resulte irracional no creerlas cuando ustedes no albergan creencia alguna sobre el asunto. En algunos contextos, una persona podría sin irracionalidad imponer criterios más estrictos de creencia que otros, y así, abstenerse de creer lo que éstos creen.

La regla 2 no manda tener una creencia como creencia racional. Si un enunciado pasa el primer test —ningún enunciado incompatible tiene un valor de credibilidad mayor—, entonces la regla 2 nos aconseja *no* creer ese enunciado si la utilidad esperada de creerlo es menor que la de no tener creencia alguna sobre el asunto. Así, la regla 2 nos dice cuándo no hay que creer un enunciado; no nos dice cuándo hay que creerlo.

Mas si un enunciado pasa el test de la regla 1, y si la utilidad (para la persona) de creer ese enunciado es mayor que la utilidad de no albergar creencia alguna sobre el asunto, entonces ¿no sería irracional para la persona no creer en el enunciado? (He ignorado hasta ahora la posibilidad de que dos enunciados incompatibles puedan empatar en credibilidad, no estando ninguno de ellos eliminado por ningún otro enunciado incompatible. Cada uno de ellos subsiste como candidato a la creencia. En tales situaciones, el cálculo de teoría de la decisión debe comparar la utilidad de creer cada uno de ellos no sólo con la de no albergar creencia alguna sobre el asunto, sino también con la utilidad de creer el otro. Si de nuevo se produjera un empate, quizá los dos valdrían.) No se gana nada abste-

---

la creencia a causa de q y de la primera regla, o será r un candidato admisible porque el q que habría de excluirlo es él mismo un candidato inadmisible a causa de p y de la aplicación de la primera regla? Lo que elimina a un enunciado merced a la primera regla ¿debe ser admisible también de acuerdo con la primera regla? Sería interesante explorar cada una de estas direcciones en el desarrollo del sistema.

¿Obliga la regla 2 a que, tras cualquier cambio de situación o de contexto, una persona tenga que recalcular la utilidad esperada de cada creencia (y de cada no creencia) que mantiene? Aquí resulta útil la distinción entre adquirir y mantener o cambiar una creencia. Manda la inercia, a no ser que haya alguna particular razón para el cambio.

niéndose de la creencia, y algo se pierde: tal es el veredicto del cálculo práctico. Si un enunciado es suficientemente creíble —esto es, no menos creíble que algún enunciado incompatible—, *¿no debería* una persona creerlo si un cálculo de teoría de la decisión le recomendara hacerlo? Podríamos transformar la más estricta versión de la regla 2 en una condición suficiente para la creencia.

> **Regla 2':** Cree (un admisible) *h* si la utilidad esperada de creer *h* es mayor que la utilidad esperada de no tener creencia alguna sobre *h*.

Sin embargo, soy reticente a aceptar esta regla sin una comprensión más detallada del modo de operar de los valores de credibilidad. ¿Puede la credibilidad de un enunciado ser mayor que la de cualquier otro enunciado incompatible, y sin embargo, ser bastante baja? (A diferencia de las probabilidades, no es necesario que las credibilidades de los enunciados excluyentes y exhaustivos sumen 1. ¿No podría, entonces, un enunciado ser más creíble que su negación y, no obstante, tener una credibilidad bastante baja?) En tal caso, no debería exigirse creer en él (en ausencia de alguna necesidad urgente de albergar alguna creencia sobre el asunto).

Esto sugiere otro requisito que puede imponerse a la creencia racional: creer un enunciado sólo si su credibilidad es suficientemente alta.[38] Cuán alto esté lo suficientemente alto variará según el tipo de enunciado de que se trate, según sea un informe observacional, un enunciado de la ciencia teórica, una creencia sobre acontecimientos históricos del pasado, etc. Así, pues,

> **Regla 3:** Cree (un admisible) *h* sólo si su valor de credibilidad es suficientemente alto dado el tipo de enunciado que es.

¿Qué fija el nivel para cada tipo de enunciado? ¿Es el nivel que maximiza la utilidad de albergar creencias de este tipo (cuando las creencias son albergadas de acuerdo con la regla 3)? Habría, entonces, dos tipos de cálculos de utilidad: el primero acerca de una creencia particular, según manda la regla 2; el segundo, acerca de un tipo de creencia para fijar el nivel de credibilidad usado en la regla 3. Pero, puesto que los enunciados pueden clasificarse de varias formas, ¿qué determina la definición y el perfil del tipo relevante? A menos que se limite de alguna manera lo que haya de contar como un tipo, el cálculo de utilidades amenaza con colapsar la regla 3 y reducirla a la regla 2.[39]

Pasar los tests de estas tres reglas —suponiendo que la tercera

regla puede definirse adecuadamente— nos da una condición suficiente para la creencia racional. Tenemos entonces la regla 4: hay que creer un enunciado si no está excluido por ninguna de estas tres reglas. (Tenemos ya: hay que creer el enunciado sólo si no está excluido por ninguna de las tres primeras reglas.) Más explícitamente,

> **Regla 4:** Cree un enunciado *h* si no hay ningún enunciado alternativo incompatible con *h* que tenga un valor de credibilidad mayor que el de *h*, y el valor de credibilidad de *h* es suficientemente alto dado el tipo de enunciado que es *h*, y la utilidad esperada de creer *h* es al menos tan grande como la utilidad esperada de no tener creencia alguna sobre *h*.

(Una vez más, una regla más estricta exigiría que la utilidad esperada de creer *h* fuera mayor que la de no tener creencia alguna sobre *h*.) La regla 2 se formuló usando el marco estándar de la teoría de la decisión; por eso habla de valor esperado. Si nuestra anterior discusión es correcta, un cálculo de teoría de la decisión debería maximizar el valor decisional de una acción, una suma ponderada de su utilidad causal, evidencial y simbólica. De manera que la segunda regla puede formularse más adecuadamente como sigue.

> **Regla 5:** Cree (un admisible) *h* sólo si el valor decisional de creer *h* es al menos tan grande como el valor decisional de no albergar creencia alguna sobre *h*.

A la hora de decidir qué creer, debemos tomar en cuenta no sólo las consecuencias causales de albergar tal creencia, sino también la utilidad simbólica de albergarla, así como aquello que indica evidencialmente el creerla. (La regla 4 debería ser correspondientemente reformulada para hablar del valor decisional en vez de la utilidad esperada.)

Ya dije antes que no es necesario que una regla para generar creencias garantice la coherencia de todas las creencias resultantes, e incluso podría llevar visiblemente a creencias incoherentes en determinadas circunstancias —en cuyo caso habrá que tomar medidas para aislar y limitar los resultados de la incoherencia—. Considérese el debatido ejemplo de Henry Kyburg sobre la «paradoja de la lotería».[40] Se va a hacer una lotería y ustedes saben que del millón de boletos de lotería, uno saldrá premiado. Para cada uno de los boletos, la probabilidad de que no resulte premiado es abrumadoramente alta. Si el que un enunciado tenga una probabilidad lo

bastante alta es suficiente para la creencia (racional), entonces ustedes no creerán que ninguno de los boletos, considerado uno a uno, vaya a salir premiado; ustedes creen que el boleto 1 no saldrá premiado, creen que el boleto 2 no saldrá premiado, ... y creen que el boleto 1.000.000 no saldrá premiado. Sin embargo, ustedes creen que se realizará el sorteo y que *alguno* de los boletos será premiado. Por lo tanto, ustedes creen que el boleto 1 será premiado, o el boleto 2 será premiado, ..., o el boleto 1.000.000 será premiado. Este conjunto de creencias es incoherente. Algunos filósofos han concluido que este ejemplo muestra que una regla que recomienda creer algo solamente porque su probabilidad rebasa cierto valor muy elevado es inadecuada. Además, si la regla añade que uno debería creer algo *sólo si* su probabilidad rebasa ese valor, entonces uno siempre creerá la conjunción de dos cosas que cree, pues esa conjunción puede tener una probabilidad menor que la de cada uno de los componentes no ciertos de la misma, y por lo tanto, en algunos casos, caer por debajo de la probabilidad mínima para la creencia.

El punto de vista que hemos presentado no hace de la probabilidad el único determinante del valor de credibilidad de un enunciado. Sin embargo, tiene que vérselas con este tipo de asuntos. Siempre que dos enunciados deban ser creídos, ¿hay que creer también su conjunción? De la estructura aquí presentada ¿puede decirse concluyentemente que resultan incoherencias? Y si es así, ¿qué reglas hay que invocar para limitar sus efectos? (Nuestro análisis puede discurrir en los términos de la regla 2 en vez de en los complicados términos de la regla 4; cualquier condición suficiente para la creencia se enfrentará a cuestiones similares.)

Alguien que aplique la regla 1 (junto con la cláusula para casos de empate) no llegará a creer dos enunciados que sabe que son incoherentes entre sí. No obstante, puesto que reconocer incoherencias no es un asunto mecánico, no hay ninguna garantía de que alguien que aplique sinceramente esa regla no se vea arrastrado (sin saberlo) a creencias incoherentes. Una vez reconocida la incoherencia, empero, la regla 1 sostiene que el enunciado con menor credibilidad no es un candidato para la creencia. Puesto que esta regla atiende a las alternativas a un enunciado que son incoherentes con él, su aplicación cuidadosa garantiza que las creencias de una persona serán coherentes por pares. La persona no albergará dos creencias, $h1$ y $h2$, que son incoherentes entre sí. Pero ¿qué ocurre con círculos más amplios de incoherencia? En la paradoja de la lotería tenemos el enunciado de que un boleto entre un millón será premiado y el enunciado individual para cada boleto de que no resultará ga-

nador. Cualquier par de ese millón y un enunciado es coherente —ambos pueden ser verdaderos a la vez—, pero no todos los enunciados que componen el millón y uno pueden ser verdaderos. Éstos forman un conjunto incoherente.

Pero no hemos exigido que el conjunto total de creencias sea coherente; sólo que las creencias sean coherentes a pares. Si ustedes desean una proporción muy alta de creencias verdaderas, crean cada uno de los enunciados del millón y uno; estarán en lo cierto un millón de veces. «Mas si el conjunto es incoherente, usted sabe inequívocamente que se equivocará una vez», objetará alguien. Es verdad, pero ¿no sería yo racional en otra circunstancia si eligiera que mis creencias se formaran mediante un proceso del que sé que me dará una creencia falsa una vez entre un millón, no como asunto de lógica —ese millón y una creencia son coherentes—, sino como asunto de hecho? ¿De qué modo han cambiado las cosas, de qué modo se ha alterado la deseabilidad de secundar el procedimiento de formación de creencias cuando pasamos a un error garantizado por la lógica —porque las creencias son incoherentes—? Es verdad que cuando yo sé que las creencias son incoherentes, haría mejor evitando que entraran todas en las premisas de un argumento que podría jugar con la incoherencia; pero éste es un asunto que tiene que ver con aislar los resultados de la incoherencia.

En esta situación, no podemos ensamblar indefinidamente creencias hasta dar con nuevas creencias. El conocimiento de que un enunciado particular —por ejemplo, una conjunción— es incoherente (dentro de la red de factores que determinan la credibilidad) entra en el valor resultante de credibilidad del enunciado para darle una credibilidad mínima. (Eso dependerá también, obviamente, de los puntos acumulados por el enunciado entrante según el cual la conjunción *es* incoherente.) La negación de la conjunción que se sabe incoherente será incompatible con ella y tendrá siempre un puntaje más alto de credibilidad. De aquí que, por la regla 1, esa conjunción incoherente no sea un candidato admisible para la creencia.

No obstante eso, ¿hasta dónde podemos llegar en la dirección de la conjunción incoherente? ¿Cuántos enunciados coherentes podríamos conjuntar y creer? En la situación de la lotería, la regla 1 (y la regla para empates) nos prohíbe desplazarnos a cualquiera de dos conjunciones distintas que sean incoherentes entre sí —por ejemplo, el enunciado de que un boleto ganará, pero que no estará entre los primeros 500.000, y el enunciado de que un boleto ganará, pero que no estará entre los segundos 500.000—. Además, en estos casos de loterías, cuanto más conjuntamos, menor es la credibilidad de

la conjunción resultante. (La probabilidad desciende, y ése es *uno* de los factores que afectan a la credibilidad resultante.) A partir de cierta masa crítica, la credibilidad de la conjunción caerá por debajo de la de algún enunciado competitivo incompatible con ella, y por lo tanto, dejará de ser un candidato admisible para la creencia. Si la situación es simétrica con respecto a las probabilidades, estaremos expuestos a un buen número de conjunciones similarmente estructuradas, cada una de las cuales será un candidato admisible de acuerdo con la regla 1. Pero la regla 2' no necesariamente acepta cualquier conjunción máxima de este tipo, aun dejando de lado la cuestión de los empates. ¿Qué beneficios se desprenderán realmente de creer en una conjunción máxima (coherente) así? (Recuérdese que alguna conjunción de menor tamaño coherente con esta conjunción máxima tendrá una credibilidad aún mayor.) Cuando aplicamos las reglas 1 y 2' en la situación de la lotería, y operamos con credibilidades (no simplemente probabilidades), podríamos esperar algo como lo siguiente.

La persona cree que uno u otro boleto resultará vencedor. Pongámosla, sin embargo, ante un determinado boleto, y la persona creerá que no resultará ganador. Pongámosla ante dos determinados boletos, y creerá que ninguno de los dos resultará ganador. Para cualquier par de boletos, siempre creerá que ninguno de los dos resultará ganador. Lo mismo por lo que hace a los tríos. En algún momento, las cosas se hacen vagas. Finalmente, para un *n*-tuplo grande, la persona dejará de albergar la creencia de que ningún boleto de ese grupo resultará ganador. No cree que ninguno de los primeros 900.000 boletos resultará ganador. Ni siquiera cree que ninguno de los primeros 490.000 ganará. ¿Dónde se planta, pues, su creencia? ¿Qué es lo que cuenta? Es posible que, en una situación particular, haya alguna razón para creer en una conjunción substancial o en otra, y en *esa* situación, el que albergue esa creencia esté determinado por la utilidad de albergarla. Pues bien; eso —en lo que hace a la última sentencia, al menos— suena muy parecido a la situación que he descrito en relación con la lotería.*

Si nuestras creencias pueden ser incoherentes —aunque las reglas operan para tratar de mantenerlas coherentes a pares—, ¿cómo

* No creo que sea una tarea especialmente urgente la reducción de la vaguedad respecto del tamaño de la conjunción aceptable más allá de la precisión alcanzada por los cálculos inspirados en la teoría de la decisión. Lo que ha interesado a la gente en la paradoja de la lotería no es el aspecto *sorites* de la misma; ¿exactamente cuántos granos de arena convierten en una duna a una duna?

tenemos que aislar el daño que esa incoherencia podría causar? Es harto sabido que de una incoherencia se puede deducir cualquier enunciado usando las reglas de inferencia estándar de la lógica. (De *p* & no-*p* podemos deducir que *p*. De *p* podemos deducir que *p* o *q* para cualquier enunciado arbitrario *q*. De *p* & no-*p* podemos deducir que no-*p*. De *p* o *q* y de no-*p* podemos deducir que *q*. Además, no es necesario que empecemos con la explícita contradicción que es la conjunción *p* & no-*p*; podemos empezar simplemente con dos enunciados distintos, el enunciado de que *p* y el enunciado de que no-*p*, y la deducción procederá como antes.)

Hay varios mecanismos que se pueden usar para evitar esa escalada en la creencia. Sugiero que para transferir legítimamente una creencia desde las premisas hasta la conclusión en una inferencia deductiva no sólo deba ser creída cada premisa, sino que deba ser creída también la conjunción de las premisas. (O que, al menos, la conjunción de las premisas no deba ser descreída.) Esto nos proporciona una sexta regla, que se aplica a las inferencias deductivas.

> **Regla 6:** Cree *q* porque se infiere de las premisas *p*1, ..., *pn* en una inferencia deductiva explícita sólo si cada una de las premisas *pi* es creída *y* sólo si su conjunción *p*1 & *p*2 & ... & *pn* es creída también.

Cuando evaluamos un enunciado como candidato a la creencia, la regla 1 nos invita a determinar si hay algún otro enunciado incompatible con el primero que tenga un valor de credibilidad mayor. No se presupone ningún umbral de credibilidad uniforme que tenga que pasar cualquier candidato admisible. El enunciado *s*1 podría ser excluido como candidato para la creencia por algún enunciado incompatible *s*2 que tiene un valor de credibilidad mayor, y sin embargo, el enunciado *s*3, que no guarda relación alguna, pero que tiene un valor de credibilidad menor que *s*1, podría ser admitido para la creencia porque no hay ningún enunciado incompatible con *él* que tenga un valor de credibilidad mayor. El umbral de credibilidad es un umbral comparativo; de aquí que cambie de un contexto a otro, en la medida en que esos contextos entrañen distintos tipos de enunciados competitivos con credibilidades diferentes.

Las reglas 2 y 5 sostienen que son las consecuencias de creer un enunciado admisible las que deben determinar si ese enunciado ha de ser creído. Cualquier teoría completa dará margen para *alguna* consideración sobre los efectos de creer *p*. Pues el hecho de creer *p* (por ciertas razones, o como resultado de cierto procedimiento) constituirá, si acontece, un hecho adicional en el mundo, y ese he-

cho puede ser significativo en sí mismo. Las probabilidades condicionales de varios enunciados pueden cambiar, incluida la del mismo enunciado *p* que nos ocupa; la probabilidad condicional de *p*, dadas las varias razones que militan a su favor y dado que ustedes creen *p* por esas razones, puede ser distinta de la probabilidad de *p* condicionada exclusivamente a esas razones. O creer *p* puede tener consecuencias causales que cambien lo que resulta verdadero en la situación que se sigue de creerlo.[41]

La regla 5 va más allá, sin embargo, y añade una consideración de la utilidad causal, evidencial y simbólica de creer *p*. No pueden esas utilidades tornar admisible una hipótesis inadmisible para la creencia racional. (Por otra parte, aunque *p* no resulte racionalmente creíble, la regla 5 no niega que creer *p* pueda ser racional en algunas circunstancias, es decir, que creer un *p* inadmisible podría traer consigo la mayor utilidad esperada o el mayor valor decisional.)[42]

Así como alguien podría evitar la investigación de determinados asuntos en una sociedad dada a causa de que sus predicciones redundarían en consecuencias sociales dañinas —alguna creencia verdadera, pero también una muchedumbre de distorsiones y malos usos—, así también alguien podría evitar creer algo por el efecto que esa creencia previsiblemente tendría en él mismo, en su carácter y en su modo de conducta. Eso no le obliga a mantener la creencia opuesta; le conmina más bien a no tenerla. Análogamente, alguien podría evitar una creencia por lo que ésta revelaría de sí mismo, aun sin causarlo, y por aquello por lo que ciertas creencias valen y simbolizan.

Creencia

¿Por qué *creer* en algo? ¿Qué hacen por nosotros las creencias? ¿A qué funciones sirven? ¿Por qué habríamos de (querer) tener creencia alguna? Porque el mundo cambia de manera irregular necesitan los organismos tener mecanismos adaptativos que respondan a circunstancias locales; ¿acaso no puede cargar con toda la tarea una estructura permanente y unas respuestas preprogramadas (como los ritmos circadianos adaptados a la estable regularidad del día y de la noche)?[43] El condicionamiento operante de la conducta confiere a los organismos cierta adaptabilidad, pero tiene dos inconvenientes: no suministra de inmediato una conducta nueva y adecuada en situaciones nuevas, o una extinción suficientemente rápida de conductas viejas, previamente reforzadas, pero ya inútiles; y no sumi-

nistra una conducta nueva y distante en el medio a menos que esa nueva conducta esté vinculada a la conducta actual por alguna cadena continua de refuerzo.[44] Las creencias son mutables, y cuando se basan en razones y en el razonamiento que arroja conclusiones nuevas, sopesando las razones a favor y las razones en contra, pueden ser ajustadas para sintonizar con situaciones nuevas o cambiantes, y así, influir útilmente en la conducta de un organismo enfrentado a tales situaciones.

Mas ¿por qué es necesario *creer* cualquier enunciado o proposición? ¿Por qué no limitarse a asignar probabilidades a todos y cada uno de los enunciados sin llegar a creer definitivamente en ninguno de ellos? ¿Por qué no actuar, en situaciones de elección, de acuerdo con esas probabilidades, (acaso) maximizando la utilidad esperada? Tal es la posición del bayesianismo radical, y tiene su atractivo.[45] El primer resultado proposicional de la estimulación sensorial podrían ser juicios (mutables) de probabilidad, sin necesidad de formular enunciados puros que representen esa estimulación sensorial como evidencia (cierta, o probable) para otros enunciados. Puesto que no estamos dispuestos a actuar de acuerdo con lo que «creemos» en todas las circunstancias, a jugarnos la vida en cada creencia particular, ¿no sería más prudente tratar todos los juicios simplemente como *grados* de creencia, como asignaciones de probabilidades a los enunciados? Si prescindimos de cualquier noción de creencia, no es necesario formular reglas de aceptación para lo que ha de ser creído —una ardua tarea—. Todo lo que se necesita son reglas para la continua modificación de las probabilidades.

Por lo demás, el coste del bayesianismo radical tampoco parece tan grande. De acuerdo con él, el científico, o la institución de la ciencia en un momento dado, no acepta o cree teorías o enunciados legaliformes; ocurre más bien, según nos enseña Carnap, que la ciencia asigna a esos enunciados grados determinados de probabilidad. Antes pensábamos que el científico creía esos enunciados, al menos *tentativamente*. Pero la filosofía de la ciencia de este siglo no ha dejado de insistir en que las teorías científicas y las leyes formuladas son, como mucho, altamente probables. ¿Cómo habría entonces de replicar a la idea, según la cual eso es todo lo que la ciencia nos *dice*? (¿Resultaría por ventura más plausible afirmar que la ciencia nos dice que determinadas cosas son definitivamente verdaderas pero que los enunciados que declaran eso son a lo sumo probablemente correctos?) Y si ustedes proclaman que *ustedes* creen definitivamente algo, que no se limitan a tener algún grado de creencia en ello, el bayesiano radical lo traducirá diciendo que ustedes confieren a eso

un grado de creencia 1: que están dispuestos a hacer apuestas infinitas, o cualquier apuesta finita, a su favor, que están dispuestos a apostar cualquier cosa por ello. (Y si ustedes no lo están, el bayesiano dirá no entender el contenido de la proclama, según la cual ustedes creen en ello y no se limitan a tener algún grado de creencia sobre ello.)

A despecho de su aparente vigor, no está claro que esta posición del bayesianismo radical pueda formularse coherentemente. Las probabilidades de los enunciados (no las meras creencias en ellos) tienen que tener aplicación en situaciones de elección, pero una situación de elección es una situación a la que una persona *cree* que está enfrentada, esto es, una situación en la que la persona cree que puede ejecutar varias acciones alternativas *A*1, ..., *An*, en la que cree que *A*1 puede tener varios resultados posibles *Oi* (dependientes seguramente del estado del mundo *Sj* que acontezca), etc. Para decirlo todo, la persona actúa de acuerdo con probabilidades, prob(*Oi*/*A*1) o prob(*Sj*/*A*1), mas esas probabilidades se dan *dentro* de una estructura de creencias sobre las situaciones de elección. ¿Es acaso posible que estas últimas no sean creencias, sino probabilidades aplicadas a enunciados? Pero el verdadero solio de las teorías de la probabilidad y de la utilidad personal, o el trasfondo interpretativo que requieren, entraña la existencia o la atribución de *creencias* a las personas cuyas elecciones se toman como indicios de preferencias o de juicios de probabilidad. Sin esas creencias acerca de la situación en la que está, sus elecciones no serían indicios de esas preferencias o juicios de probabilidad particulares. La definición teórica de estas últimas nociones presupone la atribución de ciertas creencias a la persona.

Eso vale también para el argumento bayesiano —«radical» o no— principal, según el cual los grados de probabilidad deberían satisfacer los axiomas del cálculo de probabilidades. Se trata del «argumento del libro holandés»: si los grados de creencia representan una disposición a realizar ciertas apuestas, y si los grados de creencia *no* satisfacen los axiomas de la teoría de la probabilidad, entonces la persona estará dispuesta a entrar en una serie de apuestas con las que no puede ganar ningún dinero y puede perder alguno. Mas, para que este argumento funcione, la persona debe *creer* que está frente a una situación de este tipo; si no lo cree, no apostará, o no se comportará adecuadamente. Pues las apuestas que está dispuesta a hacer dependerán de sus *creencias* acerca de la estructura de la situación (de la apuesta), de sus creencias acerca de exactamente cuántas ganancias obtendrá exactamente en qué eventualidades, etc.

Si la persona se limita meramente a pensar que es muy probable que se halle en una determinada situación de apuesta, pero que sin embargo hay alguna probabilidad de que sus acciones tengan resultados diferentes de los prescritos por la estructura de la apuesta —por ejemplo, que haya una probabilidad de 0,1 de que si ella anuncia que apostará por el enunciado *p*, los ángeles descenderán y transformarán el mundo—, entonces, en muchas de esas situaciones, la persona no se atendrá a los axiomas del cálculo de probabilidades con respecto a los enunciados que (al observador externo) parecen los únicos relevantes en la apuesta. Sólo si la persona cree que se da la situación de apuesta puede el argumento del libro holandés inferir su conclusión.

El bayesiano radical podría replicar que las referencias a creencias resultan necesarias para explicar o definir su noción de grados de creencia (probabilidad personal) y para formular su argumento del libro holandés, y que, ello no obstante, las creencias no existen. Los grados de creencia, sí: y pueden y deben ser postulados como factores explicativos a la hora de dar cuenta de la conducta de la gente. Sólo para orientarnos y para que comprendamos lo que se postula, añadirá el bayesiano, presupone él creencias, pero esa escalerilla puede tirarse después de haber culminado la subida.

El bayesiano radical se enfrenta también a complejidades prácticas formidables. La tarea de asignar probabilidades a todos y cada uno de los enunciados y combinaciones de enunciados bien formados es abrumadora. Las creencias pueden aligerar extraordinariamente esa tarea si se prescinde de asignar probabilidades a cualquier cosa que sea incompatible con las creencias de ustedes —si todas esas probabilidades son ignoradas automáticamente o reciben un valor cero—.

Isaac Levi, un crítico del bayesianismo radical, ha sostenido que las creencias funcionan como un criterio de posibilidad seria. Una vez que algo llega a ser creído, todas las posibilidades incompatibles con ese algo pueden ser ignoradas. Las creencias pueden ser reexaminadas; pero mientras se albergan, se consideran ciertas, como si no hubiera posibilidades serias de que anduvieran erradas.[46] La estructura teórica construida por Levi es impresionante en muchos sentidos, pero este tratamiento de la creencia le lleva a intrincadas dificultades. Por alguna razón, el investigador puede dejar de tener certeza en su creencia de *p*; pero mientras alberga esa creencia, tiene la certeza de la misma, y tiene que examinar razones que harían a *p* incierto mientras él mismo está creyendo *p* y, por lo tanto, creyendo que *p* no tiene posibilidades serias de ser falsa.

Levi consigue moverse por entre este campo minado, pero la trayectoria que describe es tortuosa e implausible.[47] Las minas fueron colocadas por la estructura de su propio punto de vista; a un observador externo le resulta difícil creer que la aventura era realmente necesaria.

¿Hay un camino entre los meros grados de creencia del bayesiano radical y las creencias exageradamente tenaces y obstructivas esbozadas por la teoría de Levi? He aquí una sugerencia (muy) provisional. Una creencia excluye posibilidades en un (tipo de) contexto. En otro tipo de contexto, esas mismas posibilidades no serían excluidas. Yo creo que mi colega joven recién llegado no es un pervertidor de menores. (Si se me pide una lista de personas en el edificio de filosofía que no son pervertidores de menores, no dudaré en poner su nombre en esa lista.) Pero el contexto cambia; necesito que alguien vigile a mi hijo menor por dos semanas. Un error aquí sería una cosa muy seria —lo que anda en juego se ha disparado—. Ahora pensaré con más cuidado. No es que no creyera antes en la inocencia de mi colega. En aquel contexto, y para aquellos propósitos, creía en ella; no consideré o no asigné probabilidades a la posibilidad de que fuera un pervertidor de menores. En el actual contexto, con cosas más importantes en juego, considero qué probabilidad podría haber.

El bayesiano radical dará la bienvenida a eso. «Exactamente como pensaba», dirá. «Usted nunca creyó en ello; usted se ha limitado siempre a asignarle una probabilidad». Quizá sea ésa la situación. Yo asigno probabilidades a los enunciados sólo cuando lo que anda en juego lo justifica, sólo cuando el producto de la probabilidad por la utilidad en cuestión es suficientemente grande. (Ser suficientemente grande acaso dependa no de su magnitud absoluta, sino de la proporción de la magnitud total de que podría tratarse —cinco dólares pueden marcar la diferencia entre pedir uno u otro plato en un restaurante, pero no entre comprar uno u otro coche—.) Procedo primero a una estimación *aproximada* de si la magnitud *podría* ser suficientemente grande, y sólo si es éste el caso, asigno realmente una probabilidad al acontecimiento o al enunciado. (Ya hemos argüido, sin embargo, que el bayesiano radical no puede decir esto acerca de *cualquier* creencia.)

Quizá haya rasgos contextuales relevantes distintos de la utilidad que anda en juego. Ciertas empresas, como las editoriales científicas o de libros de texto, pueden introducir criterios profesionales que debe satisfacer un texto para ser publicado. Quizás estas normas intelectuales más estrictas se basan a fin de cuentas en al-

gún argumento general acerca de lo que podría estar en juego —ustedes no saben quiénes actuarán basándose en la información por ustedes publicada, ni en qué circunstancias—. O quizá contribuyen esas normas a definir la naturaleza de la empresa, revelándonos información sobre la misma.

En lo atinente a las creencias cotidianas, la regla 1 proporciona un criterio suficiente de credibilidad, a saber: que nada incompatible es más creíble. (Eso no es suficiente para la creencia, sin embargo; al menos hay que pasar también el test de la regla 3.) Para la aceptación en ciencia, no obstante, se aplica un criterio de credibilidad más restricto: se puede exigir a un enunciado que alcance cierto nivel de credibilidad, no simplemente que sea más creíble que cualquier alternativa incompatible. ¿O acaso deberíamos limitarnos a decir que esta situación quedará fijada por la regla 3, aplicada a contextos científicos? La regla 3 misma es contextual. Los criterios científicos, una vez conocidos, tienden a extenderse a —algunos dirían a «invadir»— otros contextos, primero a contextos públicos relacionados con asuntos de política social, y luego incluso a contextos interpersonales o personales.

Sólo recientemente ha resultado claro que, a la hora de hacer afirmaciones sobre la eficacia de los medicamentos, hay que asegurarse de que esas afirmaciones estén respaldadas por experimentos con un mecanismo de doble ceguera; si el paciente o el juez de la (supuesta) mejora sabe si la medicación ha sido administrada, eso puede contaminar los resultados. La experimentación doblemente ciega constituye hoy un criterio para la estimación de razones y evidencias en esa área. (Tal cambio es una aplicación de un criterio constante, según creo, que ordena controlar todas las variables relevantes.) El criterio más exigente no resulta apropiado en cualquier área: yo puedo sostener razonablemente un enunciado causal cotidiano sin necesidad de haber realizado un experimento doblemente ciego que lo respalde.

Cuán estrictos tengan que ser los criterios para estimar razones dependerá de lo que ande en juego, de cuán importante o serio pueda ser un error, de cuánta energía, cuánto tiempo y cuántos recursos habría que dedicar al empleo de procedimientos que satisfacieran esos criterios más estrictos, de la naturaleza general de la empresa, etc. Si hay distintos procedimientos *P*1, ..., *Pn* que satisfagan los criterios de distintos grados de rigurosidad, podemos abordar la cuestión misma de qué procedimientos usar como un problema de decisión, calculando los costes y los beneficios de cada procedimiento en su contexto específico.[48] Pero ¿qué criterios o pro-

cedimientos usar para realizar ese cálculo? ¿Estamos condenados a una circularidad tan vitanda como ineludible? Para responder a la cuestión teórica de cuáles son los mejores procedimientos y en qué circunstancias, me parece que lo que queremos es usar el procedimiento más estricto que satisfaga los criterios más estrictos con objeto de disponer de una respuesta que sirva de-una-vez-por-todas y que no dependa de las exigencias de nuestra particular situación de ahora mismo. (Eso es lo que distingue a este contexto teórico.)

No sólo la creencia anda vinculada al contexto; también la racionalidad. Llamar a algo racional es proceder a una *evaluación:* sus razones son *buenas* razones (de un cierto tipo), y satisface los criterios (de un cierto tipo) que *debería* satisfacer. Esos criterios, ya va dicho, pueden variar de un ámbito a otro, de contexto a contexto, de ocasión en ocasión. Por eso deberíamos ser cautelosos al concluir que alguien es irracional simplemente porque sus razones no satisfacen los criterios más estrictos que somos capaces de formular. Puede que satisfagan los criterios adecuados a su contexto, los criterios más estrictos que la teoría recomendaría en tal situación. Podría ser irracional para él, diría la más estricta teoría, atenerse aquí a criterios más estrictos.

Si algo es plenamente racional cuando satisface todos los criterios (de un cierto tipo) que debería satisfacer en relación con las razones, también puede haber aquí una noción de grado, una noción que hable de grados de racionalidad o de racionalidad en ciertos aspectos, cuando alguien satisface ciertos criterios, pero no todos, o satisface algunos hasta cierto grado, pero no completamente.

En determinados contextos, determinadas cosas se toman como dadas. Esas cosas fijan el marco en el que una persona actúa o elige, el marco en el que trata de maximizar alguna función o de que su acción revele determinada propiedad. El que una persona tome como dado algún enunciado *q* en el contexto *C* monta tanto como atenerse a él cuando está en *C* y trata de maximizar alguna función. No hace cálculos para llegar *a q;* trata de viajar *partiendo de q* a algún otro sitio. En el contexto *C* tomamos como dado *q*, y en este contexto llegamos a la creencia *r*, y ahora podemos considerar*la* como dada, pero sólo en contextos como *C* en los que resulte adecuado tomar como dada a *q*. Puesto que *q* estaba vinculada a *C*, *r* no flota libre independientemente de lo que se consideró como dado en *C* para alcanzar *r*.[49]

Las creencias están vinculadas a contextos en los que las posibilidades incompatibles con ellas son excluidas o tratadas como indignas de consideración: llamemos a este punto de vista «contex-

tualismo radical».[50] (Podemos dejar abierta la cuestión de si hay que tomar como dada o no alguna creencia en todos los contextos posibles.) Podemos identificar las creencias más plenamente con un par [*bi, Cj*]. Ese par dice que es verdad para mí que yo tomaré como dada *bi* cuando (yo crea que) estoy en *Cj*. Yo ahora soy el tal para el que esto es verdad, aunque yo podría cambiar de tal forma que esto dejara de ser verdad. El par señala una disposición que yo poseo ahora.*

Las creencias afectan a las acciones en la medida en que incorporan expectativas acerca de los resultados que esas acciones tendrán o tendrían. Después de hecha una acción, esas expectativas se cumplirán o no (o parecerán cumplirse o no) en varios grados, y esos resultados modificarán entonces las creencias que incorporan esas expectativas. Las creencias acerca del mundo se progrealimentan con acciones, y los resultados (percibidos) de esas acciones, junto con otros hechos percibidos, retroalimentan, positiva o negativamente, las creencias. El bayesiano acepta también esa retroalimentación de las probabilidades que lleva a su revisión, quizá de acuerdo con la condicionalización. Ello no obstante, esa retroalimentación marca también una distinción entre el contextualismo radical y el bayesianismo radical. El contextualismo radical ignora ciertas posibilidades en ciertos contextos. El bayesiano radical identifica esas posibilidades con una probabilidad que, cuando se multiplica con lo que anda en juego, arroja un producto que, dado el resto de magnitudes implicadas, es demasiado pequeño para afectar a la decisión —por eso (según sostiene) él también puede ignorarlas justificadamente—. En la etapa bayesiana de retroalimentación, empero, parece que esas probabilidades necesitarán revisión y «actualización» junto con todas las demás, y eso implica un esfuerzo computacional significativo. El contextualista radical, por otro lado, continúa ignorando esas posibilidades, ahora lo mismo que antes, al menos mientras siga en este contexto.[51] (En otro contexto, podría resultar necesario considerar aquellas posibilidades y sus probabilidades.)

¿Hay principios generales acerca de qué tipos de creencia resul-

* Obsérvese que el contextualismo radical, a diferencia de la sociología del conocimiento, no se socava a sí mismo. Sea *CR* la doctrina del contextualismo radical; todas las creencias se tienen en un contexto. Cuando alguien cree *CR*, también él está en un contexto, llamémosle *Cj*, y en este contexto toma como dado *CR*, excluye las posibilidades incompatibles con él, etc. Un enunciado que declare que *S* es válido exactamente en los contextos *Ci* anda él mismo a salvo, siempre que se mantenga en *Ci*.

ta adecuado tomar como dados y en qué contextos? ¿Pero en qué contexto se formulan esos principios, y qué se toma como dado aquí? ¿Tendrían tales principios una justificación racional? ¿O se trataría acaso de principios que habrían sido inculcados en nuestros ancestros y habrían conseguido funcionar lo suficientemente bien como para haber pasado hasta nosotros? ¿Podríamos esperar que la evolución hubiera inculcado un mecanismo para la creencia contextual, o más bien un mecanismo para la creencia no contextual que casara con una gran variedad de contextos lo suficientemente bien como para haber sido seleccionado? Comoquiera que se desarrolle la teoría de la creencia contextual, lo cierto es que cubre un fenómeno de creencia, por lo que seguirá habiendo espacio abierto para plantear cuestiones sobre la racionalidad de la creencia. Apuesto a que el bayesiano radical no conseguirá hacer desaparecer este tópico.

Sesgos

Hemos dicho que una persona racional no se limita a extrapolar a partir del balance neto de todas las razones que tiene hasta alcanzar una conclusión acerca de la verdad. Considerará la posibilidad de que las razones de que se percata no constituyan una muestra representativa de todas las razones existentes. Una persona racional, pues, estará autoconscientemente alerta sobre los posibles sesgos de su propio funcionamiento intelectual y de la información que recibe.

Amos Tversky y Daniel Kahneman discuten el modo en que la gente estima a veces la frecuencia de una clase por la facilidad con que ejemplificaciones de la misma pueden ser mentalmente representadas.[52] Por ejemplo, la gente sospecha que, en una muestra al azar procedente de un texto en inglés, habrá más palabras que tengan la «r» como primera que como tercera letra, porque las palabras vienen más fácilmente a la cabeza partiendo de su primera letra. Sin embargo, la frecuencia de las palabras cuya tercera letra es «r» en la muestra de palabras en las que piensa la gente no es representativa de su frecuencia en la población total de palabras. Hay un sesgo en el modo en que la información viene a la cabeza.

Los psicólogos han notado también que, al evaluar o al dar apoyo a una creencia, no usamos o sacamos de la memoria una muestra aleatoria o representativa de la evidencia relevante que poseemos. La evidencia que hemos encontrado más recientemente o más sorprendentemente pesará más en nosotros. También puede ocurrir

que, cuando conseguimos un nuevo bit de información que tiende a dar apoyo a una hipótesis o a una creencia en particular, saca de la memoria información convergente que se compadece bien con la nueva y viene en apoyo de la misma hipótesis. (La evidencia presente que apunta a una falta de agudeza visual les recuerda a ustedes aquellas ocasiones pasadas en las que ustedes no consiguieron ver las cosas demasiado bien.) Esa información encaja entonces en una pauta, en un edificio de apoyo, de manera que aun si un puntal desaparece —incluso el puntal que originalmente dio pie a los demás—, los puntales restantes son suficientes para apoyar y mantener la creencia o el punto de vista. Acaso este fenómeno ande por detrás del papel desempeñado por las mentiras y las calumnias en la vida política. No sólo no consigue la verdad ponerse al día con la información falsa —algo que ya sabíamos—, sino que incluso cuando lo consigue y se elimina un particular bit de información, los efectos producidos por la circulación de esa información *no* son eliminados, sino que continúan. (Un candidato político podría servirse de la calumnia consciente contra un oponente, confiando en que cuando éste la replique y aun la repudie, ésta habrá hecho ya su trabajo.)

Esta explicación de por qué las creencias no regresan a su estado previo cuando se revela la falsedad de alguna información en la que se fundan[53] tiene implicaciones también para los efectos que habrá de tener la información *verdadera* que apoya a una hipótesis. También ésta conseguirá sacar de la memoria otra información convergente que da apoyo a la hipótesis, mientras que no sacará paralelamente información que tienda a contar en contra de la hipótesis. De manera que, a falta de procedimientos específicamente diseñados para evocar y considerar evidencia contraria, habrá una tendencia a sobreestimar la verosimilitud o el apoyo con que cuenta una hipótesis. La evidencia en la que basamos nuestras creencias no es (en general) una muestra aleatoria de la evidencia relevante accesible a nosotros o de la evidencia que (en algún sentido) poseemos ya. Una presentación llamativa y destacada de *alguna* evidencia producirá sesgos en la evocación de otra evidencia, y así, sesgos en las creencias resultantes.[54] De aquí que, a la hora de evaluar una posible creencia, resulte de especial importancia no sólo considerar la evidencia favorable y la evidencia contraria a lo que habíamos pensado, sino hacer también esfuerzos particulares y sistemáticos para evocar toda la evidencia relevante, a favor y en contra, que tengamos.[55]

Más allá de la posibilidad de que la información que nos venga a la cabeza no sea una muestra representativa de la información que

ya tenemos, la información que tenemos quizá no sea una muestra representativa de la información que hay. La relativa frecuencia con que cierta clase de información viene a nuestra atención a través de la prensa o de las pantallas, por ejemplo, quizá no sea representativa de toda la información existente sobre el asunto. Los sesgos políticos o ideológicos en la fuente que son los noticiarios pueden hacer menos probable que información contraria al sesgo sea presentada, y ciertos tipos de información quizá resulten lisa y llanamente más difíciles de cubrir o no sean publicables porque los editores piensan que al público no le parecerá interesante o creíble esa información. Un sesgo en el tipo de información (a favor o en contra de una conclusión acerca de un asunto particular) que consigue llegar hasta nosotros a través de nuestras fuentes se representa no en las probabilidades previas de que esa información sea correcta, sino en las probabilidades *condicionales* de que este tipo de información sea ofrecida por la fuente, *dado* que sea correcta. Lo que deberíamos concluir respecto de la verdad sobre la base de la información que tenemos depende en parte de qué información diferente habría llegado hasta nosotros (a través de nuestras fuentes de información y de razones) si el balance neto entre todas las razones que hay fuera harto distinto. (El marco bayesiano parece un marco apto para representar estos asuntos.) Al pensar en los posibles sesgos de las fuentes humanas, tenemos que considerar las motivaciones y los incentivos de esas fuentes.*

* Algunos autores de libros de física nos dicen que la física apoya y da fundamento a una visión espiritual del universo, mas ¿qué ha de pensar el lego si esos autores mismos se muestran sedientos de lecciones espirituales? ¿Hasta qué punto resultan sus opiniones de lo que los hechos más plausiblemente revelan y hasta qué punto expresan sus propios deseos y anhelos? Lo que sí resultaría interesante es que algún físico nos dijera *desolado* que, a pesar de lo que él deseaba que fuera el caso, en contra de sus propias preconcepciones materialistas personales, se ha visto forzado a concluir que la física contemporánea apunta a que el universo tiene una base espiritual. (Hasta donde yo sé, esto todavía no ha ocurrido.)

Consideremos los *Consumer Reports*, que ponen en guardia respecto de las posibilidades de sesgos, o de indicios de ellos, rechazando la aceptación de publicidad y pleiteando contra cualquier empresa que cite sus evaluaciones favorables. De aquí, o eso parece, que los lectores y suscriptores puedan confiar en los incentivos de esos informes para proporcionar informaciones y evaluaciones precisas y no sesgadas. Pero ¿acaso su único incentivo es el de servir a sus lectores, o tienen también un incentivo en *complacer* a esos lectores para que renueven la suscripción? La información será valiosa para los lectores si se limita a confirmar lo que ellos ya piensan. Por eso los *Consumer Reports* tienen que informar con frecuencia de que lo que generalmente se cree *no* es así, que lo que generalmente se pensaba que era el mejor producto, de hecho, no lo es, que el producto más caro es peor que otro. ¿Qué

La racionalidad en la creencia y en la acción depende de que haya algún tipo de autoconsciencia a la hora de juzgar el proceso por el cual llegamos a tener nuestras razones. Una persona racional usará *algunos* procedimientos con los que operar y corregir otros procedimientos, para corregir sesgos en el muestreo y evaluación de

---

debe entonces pensar el lector cuando lee una historia como ésta: que bien a menudo las creencias populares andan erradas, de manera que la revista no necesita distorsionar, sesgar o enmascarar nada para seguir complaciendo a sus lectores, o que se trata de un caso en el que la revista está haciendo lo que debe hacer de vez en cuando para poder sobrevivir? Yo tiendo a la primera hipótesis, pero resulta iluminador ver el margen abierto para la segunda. Al evaluar el periodismo impreso y televisivo, tenemos que considerar también los incentivos de quienes lo producen, cómo se promocionan profesionalmente descubriendo y presentando ciertas historias —por ejemplo, sobre las vidas personales de los candidatos políticos— y cómo todo esto podría sesgar las noticias que ofrecen y la importancia conferida a esas noticias, y cómo eso podría sesgar ulteriormente los resultados de las decisiones públicas.

Las campañas políticas en los Estados Unidos hacen ahora un amplio uso de entrevistas centradas en grupos de personas minuciosamente seleccionadas para descubrir qué virtudes de un candidato, o qué defectos de un oponente, tienen un gran peso emocional para esos votantes. Véase Elizabeth Kolbert, «Test-Marketing a President: How Focus Groups Pervade Campaign Politics», *New York Times Magazine*, 30 de agosto de 1992, págs. 18-21, 60, 68, 72. Las consignas y la publicidad política usan entonces selectivamente esa información para asegurarse la victoria electoral —el cumplimiento real en el cargo del vencedor parece verse afectado muy poco—. Los temas que se ponen de relieve no constituyen una muestra aleatoria o representativa de los temas significativos descubiertos. Si esas entrevistas realizadas por los colaboradores de un candidato descubren que éste tiene una virtud y veinte defectos, mientras que su oponente tiene veinte virtudes y un defecto, la publicidad (emocionalmente poderosa y atractiva) se centrará en esa virtud y en el único defecto de su oponente. Aparentemente, muchos votantes no descuentan esa falta de representatividad, y los mismos participantes en esos grupos de entrevistados no parecen generar aversión alguna a que se exploten tan selectivamente sus inquietudes. El significativo efecto que esto tiene ahora en las elecciones constituye un problema público.

Alguien que va a comprar un libro, leyendo las citas que ofrece su faja o la publicidad sobre el mismo, puede presumir que el editor no nos está ofreciendo una muestra representativa de opiniones acerca del libro, sino que está más bien destacando las *mejores* opiniones vertidas sobre él, prestando alguna atención a fuentes que el lector podría esperablemente conocer y respetar. De manera que el futuro comprador del libro debería corregir ese sesgo de muestreo y pensar: puesto que éstas son las mejores opiniones seleccionadas de una distribución de todas las opiniones sobre este libro, mi propia opinión probablemente sea *menos* favorable que las que se me ofrecen aquí. (¿Publican los editores de libros de bolsillo páginas y páginas de recensiones repetitivamente favorables procedentes de fuentes diversas para convencer al lector de que esas opiniones están tan extendidas que puede confiar razonablemente en que su propia opinión coincidirá con ellas?

razones y en el procedimiento de estimación basado en ellos.[56] (¿Podría haber sesgos en este segundo nivel de procedimientos correctores?) Una educación universitaria no sólo debería enseñar (las técnicas de adquisición de) nuevas ideas y viejas ideas dignas de ser transmitidas, sino que debería alertarnos también sobre determinadas fuentes de sesgos informativos y evaluativos e impartir técnicas de compensación y corrección de esos sesgos.

Vale la pena decir algo sobre la noción general del sesgo. Podemos distinguir dos tipos de sesgos. El primero tienen que ver con la aplicación descompensada de criterios existentes. La discriminación en el ámbito social, por ejemplo, tiene que ver con la aplicación descompensada de criterios para diferentes grupos de individuos. No obstante, hay muchos criterios diferentes que podrían considerarse de un modo u otro relevantes para una decisión o para un tratamiento de un caso. ¿Qué determina cuál de los posibles criterios constituirá el criterio para la elección, y qué determina los pesos que se le conferirán a este último cuando los criterios considerados no constituyan una muestra aleatoria de todos los posibles criterios relevantes?

Habrá un sesgo en la selección de criterios cuando la explicación de por qué se consideran *esos* criterios y no otros, o de por qué se confieren esos pesos y no otros, entrañe en parte la creencia por parte de algunos de que esos mismos criterios y pesos deberían redundar en la exclusión o en detrimento de grupos particulares, y esa creencia sea lo que les motivó a avanzar esos determinados cri-

---

Hablando de libros, no puedo resistirme a dar noticia de otro fenómeno, un fenómeno que en realidad puede vincularse al tema presente. La prensa popular realza mucho los premios y distinciones que traen consigo un gran beneficio financiero, pero hay un premio literario crematísticamente muy modesto que recibe una atención muy significativa: el Premio Pulitzer, que se concede a la ficción, a la historia, a la biografía, a la poesía y al ensayo general. Una explicación de esa prominencia podría ser la antigüedad y el prestigio de ese premio, pero yo sugiero otra. El Premio Pulitzer se concede también a periodistas y a periódicos, lo que hace que los periódicos den a esos premios un tratamiento informativo tan destacado que merece la primera plana. Si alguien fundara un premio modesto para pintores y coreógrafos, digamos, yo le recomendaría que añadiera premios para los presentadores de noticiarios televisivos y para los productores de tales programas.

La conexión que prometí es ésta: no se puede juzgar la importancia de una historia por el relieve concedido a su tratamiento sin considerar los incentivos de quienes dan tratamiento a la historia. Mi propia concepción es que debería darse mucho relieve a *todos* los premios literarios, científicos, artísticos e intelectuales; pero cuando evalúen esta concepción, espero que los lectores no dejen de considerar también aquí la fuente.

terios. Esos criterios fueron escogidos *para* excluir ciertos casos. Llamemos a eso un *sesgo de segundo nivel.*[57] Cuando el origen del criterio es un sesgo de segundo nivel, pero pueden movilizarse otras razones en favor suyo, y el criterio es secundado en parte por esas otras razones, la situación definicional y normativa se hace más complicada. (También podemos preguntar por qué esas otras razones llegaron súbitamente a destacar, por qué consiguieron el peso que consiguieron. Si de todas las posibles razones éstas hubieran sido avanzadas *para* justificar la discriminación de segundo nivel, entonces esto constituiría a su vez una discriminación de un nivel más elevado, pero es mejor considerarla también como de segundo nivel.) Observadores muy refinados pueden a veces discutir con pericia el asunto de la discriminación de primer nivel sin atender a la posibilidad de la discriminación de segundo nivel. Un estudio frecuentemente citado por los estadísticos investiga si durante un determinado período la escuela de graduados de la Universidad de California en Berkeley discriminó a las solicitantes femeninas en su proceso de admisión.[58] Las mujeres que solicitaban plaza parecían tan cualificadas como los varones solicitantes según los criterios usados (estudios de secundaria, número de cursos recibidos sobre la materia, puntos conseguidos en el examen de graduados, etc.), y sin embargo sólo un porcentaje muy inferior de mujeres eran admitidas a la escuela de graduados. ¿No debía ser éste un caso de discriminación? No (respondieron los estadísticos), pues cuando nos fijamos en las admisiones departamento a departamento, hallamos que cada departamento admitió aproximadamente el mismo porcentaje de mujeres que de varones. Ningún departamento practicó discriminación. ¿Cómo entonces llegaron a ser tan distintos los porcentajes globales? Las solicitudes de los varones y de las mujeres no se concentraban en los mismos departamentos. Algunos departamentos admitieron un porcentaje menor de solicitantes que otros departamentos, y las mujeres presentaron más solicitudes en esos departamentos. (A modo de ilustración: si la mayoría de las mujeres presentaran solicitudes a departamentos que admitieran sólo al 10 por ciento de sus solicitantes, y la mayoría de los varones presentaran solicitudes a departamentos que admitieran al 50 por ciento de los solicitantes, entonces la escuela de graduados admitiría globalmente diferentes porcentajes de hombres y mujeres solicitantes aun cuando cada departamento admitiera el mismo porcentaje de varones y mujeres solicitantes.) Caso cerrado (dijeron los estadísticos): no hay discriminación.

Correcto. Pero todo lo que el estudio mostró es que no había dis-

criminación de primer nivel; la evidencia no consiguió probarla. Ello no obstante, aún podemos preguntar por qué diferentes departamentos admitían diferentes porcentajes de solicitantes. Presumiblemente porque la *ratio* entre los solicitantes y las posibles plazas de admisión difería de departamento a departamento. Ocurrió simplemente que las mujeres presentaron solicitudes en departamentos con un número menor de plazas de admisión por solicitante. Pero esto ¿«ocurrió simplemente»? ¿Por qué las plazas de los departamentos no son proporcionales al número de solicitantes cualificados? ¿Qué determina la dimensión de un departamento, qué determina la cantidad de recursos de la universidad destinados a posiciones docentes, a becarios, etc., en cada departamento? Seguramente muchos factores. *Supongamos,* sin embargo, que algunos departamentos están subdotados porque tienen mayoritariamente una población estudiantil femenina. Supongamos que por esa razón no están tan considerados por los administradores de la universidad o por la sociedad que destina fondos para los programas de graduados. O supongamos que otros programas de graduados estuvieran mejor dotados y fueran capaces de admitir un porcentaje de solicitudes más elevado porque su población fuera mayoritariamente masculina y eso les acarreara una mejor consideración. Algunos departamentos son entonces más pequeños que lo que podrían ser en otras circunstancias *debido* a su alto porcentaje de solicitudes femeninas; o bien, otros departamentos son más grandes debido a su alto porcentaje de solicitantes masculinos. No estoy diciendo que esto sea así; me limito a describir una posibilidad no completamente implausible, caso de darse la cual las estadísticas de Berkeley *casarían* con una pauta de discriminación de segundo nivel. Las estadísticas por sí solas no conseguirían demostrar eso; demostrarlo requeriría investigar una cuestión estructural acerca de la organización de la universidad, a saber: por qué difieren los varios departamentos en el porcentaje de solicitantes que pueden admitir.[59]

En situaciones de discriminación de segundo nivel, los criterios aplicados (o los pesos que se les confieren) no son una muestra aleatoria del conjunto de posibles criterios relevantes. Además, cualesquiera razones que traten de justificar esos particulares criterios (o pesos) no son una muestra aleatoria (o un subconjunto objetivamente justificado) de las razones evaluativas posiblemente relevantes. El muestreo está aquí intencionalmente sesgado. En otros casos, el asunto puede resultar menos claro. Considérese el debate entre críticos literarios acerca de la naturaleza del canon literario y de las condiciones para la inclusión en él. Podría haber una discrimi-

nación de primer nivel de las mujeres, de los escritores minoritarios y de los escritores pertenecientes a otras culturas si alguno o muchos de ellos fueran excluidos del canon a pesar de satisfacer los mismos criterios que han permitido la inclusión de una mayoría de escritores varones. Con todo, aun si los criterios existentes se aplican correctamente, también podemos investigar la posibilidad de una discriminación de segundo nivel en el establecimiento del canon. ¿Por qué hay que aplicar precisamente *estos* criterios? ¿Hay otras virtudes o razones que hagan a las obras dignas de ser estudiadas con igual seriedad con métodos idénticos o muy similares? ¿Hay otros métodos interesantes y fértiles que podrían divisarse para estudiar e iluminar esas obras diferentes? Eso no significa que uno sea capaz de decir cuáles son esos criterios desde el comienzo. Los griegos sabían que los dramas de Esquilo y de Sófocles eran acreedores de una atención sostenida antes de que Aristóteles escribiera su *Poética* para formular explícitamente los criterios que esos dramas satisfacen.[60] Buscar nuevos criterios no significa denigrar las virtudes captadas por los existentes.

Creo que donde la sociología del conocimiento tiene algo importante que decir es en el ámbito de la selección entre varios posibles criterios. Hay muchos posibles tipos de razones a favor y en contra de cualquier creencia —y desde luego, de cualquier creencia que tenga que ver con cuestiones sociales o normativas controvertidas—, y hay muchos posibles criterios para evaluar esas razones. Nadie busca todas las posibles razones imparcialmente y les confiere igual peso —incluidos aquellos de nosotros que hacemos algún esfuerzo por examinar las razones en contra— y las razones que se aceptan no son una muestra aleatoria de todas las posibles razones. Parece razonable pensar que los factores estudiados por los sociólogos clásicos del conocimiento (en la tradición de Karl Mannheim), tales como la posición de clase, el nivel educativo, la red de vínculos de grupo, etc., influirá en la elección por parte de la persona entre las varias razones y los varios criterios de evaluación posibles, en lo destacados que le parezcan algunos y en el peso que les confiera. (Recientemente, otros autores han insistido en el género, en la raza y en la preferencia sexual.) Enfrentada a una situación social complicada, gente en distintas posiciones sociales puede atender a y centrarse en aspectos diferentes y, consiguientemente, apelar a diferentes principios (adecuados para esos aspectos) que llevan a diferentes conclusiones. Todos pueden estar en lo cierto: aquello a lo que cada uno ha atendido *es* una razón para sus respectivas conclusiones (opuestas). Todos pueden ser también racionales: al creer cosas por

razones, al llegar a conclusiones basándose en la evidencia, al apelar a criterios para la creencia y la evaluación que tienen mucho en su favor. Sin embargo, todos creen cosas por razones que son sólo algunas de las razones, llegan a conclusiones basándose sólo en alguna evidencia, apelan sólo a algunos criterios que tienen mucho en su favor. Los factores sociales (estudiados por los sociólogos) actúan restringiendo la gama de posibles consideraciones relevantes que se consideran. Dentro de esa gama, nuestras creencias y nuestras acciones pueden ser racionales. Y aun sin hacerlas relativas a esa gama, no son completamente irracionales —dado que hay *alguna* razón para ellas: son al menos *prima facie* racionales—.

En la introducción atendimos de refilón a la tesis de que la racionalidad misma está sesgada. En vez de ello podría sostenerse que, aun cuando la racionalidad no está objetablemente sesgada en el propósito de los objetivos que persigue —corrige este sesgo lo mejor que puede—, está sesgada en los objetivos mismos que persigue y en la manera de perseguirlos. ¿No excluye acaso la racionalidad a la emoción, a la pasión y a la espontaneidad, y no son éstos componentes de la vida? Pero la racionalidad puede andar tras ellos. Incluso la racionalidad de la teoría de la decisión puede recomendar la toma de muchas decisiones sin pensamiento ni cálculo si el hacerlo resulta más valioso que las pérdidas en que podrían incurrir esas decisiones menos meditadas, o si el proceso de cálculo mismo interfiriera en la naturaleza de otras relaciones valiosas, como el amor o la confianza.[61] Es verdad que si la racionalidad fuera a recomendar todo eso, lo haría fundándose en razones que considera y evalúa como buenas razones, de manera que *esta* recomendación no sería irreflexiva. Pero esto no es lo mismo que *calcular.* Ser sensible a las razones no implica la explícita consideración de las mismas. La racionalidad puede ser modesta y elegir mantenerse al margen algunas veces, o incluso, en determinados tipos de circunstancias, casi siempre.

CAPÍTULO 4

RAZONES EVOLUCIONARIAS

La racionalidad de una creencia o de una acción es una cuestión de sensibilidad respecto de las razones a favor y en contra y del proceso que genera esas razones. ¿Por qué tiene que ver la racionalidad con esas razones? He aquí una respuesta: las creencias y las acciones tienen que tener determinadas propiedades (como la verdad o la satisfacción de un deseo), y es más probable que las tengan si son sensibles a todas las razones a favor y en contra. (¿Podría por ventura otro proceso que no tuviera nada que ver con la consideración o con la ponderación de razones ser aún más fiable en punto a alcanzar el objetivo?) Sea o no la consideración de razones el método más efectivo o más fiable para lograr nuestros objetivos cognitivos, ¿por qué ese método resulta de alguna manera efectivo? ¿Qué conexión hay entre las razones a favor y en contra y esos objetivos? ¿Qué es lo que hace que algo sea una razón? En el torbellino de información existente, ¿qué constituye a algo en razón a favor o en contra de una creencia o de una acción?

Venimos a una creencia a través de algún proceso general para llegar a, para mantener y para revisar creencias. Para distintos tipos de creencias, o incluso para el mismo tipo en distintas ocasiones, podemos emplear procesos distintos. Seguir un proceso particular en una particular ocasión puede llevarnos a una creencia que es verdadera; usar ese proceso puede causar en nosotros (invariable o probabilísticamente) creer la verdad. La utilización de razones puede desempeñar un papel en el hecho de que un proceso sea una causa probabilística de creer la verdad, y las diferencias entre procedimientos de razonamiento pueden afectar a la eficiencia o a la eficacia del proceso. Resulta plausible decir que creer *h* por la razón *r* nos conducirá a lograr nuestro objetivo cognitivo de la creencia verdadera sólo si hay alguna conexión entre la verdad de *r* y la verdad de *h*. Es esta conexión entre las razones y la verdad de aquello para lo que son razones lo que explica el vínculo que se da entre creer por razones y creer la verdad. ¿Cuál es entonces la naturaleza de la conexión entre las razones y aquello para lo que son razones?

RAZONES Y HECHOS

Respecto de las razones para la creencia, la literatura filosófica registra dos puntos de vista. Uno, el punto de vista *a priori*, sostiene que una razón *r* para una hipótesis *h* se halla en alguna relación *R* con *h*, tal que la facultad de la razón puede aprehender que esta relación (¿estructural?) es una relación de apoyo. Las razones son cosas que la mente tiene la capacidad de reconocer.[1] El problema es éste: ¿por qué esperar que *h* sea realmente verdadera cuando *r* es verdadera y *r* se halla en la relación *R* con *h*? Si se replica que lo esperamos porque *r* es una razón para *h*, aún podemos preguntar qué explicación se ofrece de por qué *h* es (a menudo) realmente verdadera cuando *r* lo es.[2]

El segundo punto de vista, el punto de vista fáctico, sostiene que *r* constituye una evidencia para *h* cuando se halla en una cierta relación contingente fáctica con *h*. En *Philosophical Explanations* yo he sostenido que la conexión evidencial es una relación fáctica y he presentado una noción de ella en términos de la relación de rastreo entre evidencia e hipótesis (y en términos de aproximaciones probabilistas a esa relación).[3] Pero no deseo insistir en esa particular noción de la relación fáctica.

Cuando se define adecuadamente la relación fáctica, en el bien entendido de que se da, constituirá, en combinación con *r*, una razón para creer *h*. Mas sin el conocimiento de que tal conexión se da —aunque se dé— y sin ninguna conexión estructural evidente entre *r* y *h*, ¿constituirá *r* una razón para creer *h*? El punto de vista fáctico parece dejar de lado lo que más llama la atención del punto de vista *a priori*, a saber: que en los casos particulares la conexión de la razón parece (casi) evidente. (Obsérvese que la noción de *evidencia* podría casar con un enfoque puramente fáctico aun si no lo hiciera la noción de *razón*.)

Sugiero una combinación de los dos puntos de vista. Una razón *r* para *h* es algo que se halla en una cierta —dejemos que una ulterior pieza de la teoría concrete eso— conexión fáctica con *h*, mientras que los contenidos de *r* y de *h* se hallan en una cierta conexión estructural que aparece llamativamente ante nosotros haciendo a *h* (más) creíble dado *r*. La relación de razón es una conexión fáctica que aparece, independientemente de la experiencia, como una conexión de apoyo. (Pueden ustedes substituir el término *apoyo* por alguna descripción que prefieran los propugnadores de la facultad de la razón.)

Podría haber distintas bases de partida para nuestra actuación

de acuerdo con una conexión fáctica: la acción podría estar preprogramada (si la conexión fáctica se dio en el pasado y se produjo una selección evolucionaria para esta secuencia hecho-acción), o la acción podría resultar del condicionamiento operante. Hay una tercera base. Actuar según *razones* entraña *reconocer* una conexión de relación estructural entre contenidos. Tal conexión hubiera podido ser ella misma útil y resultar positivamente seleccionada. La propiedad de que una cierta conexión fáctica nos *parezca* a nosotros evidentemente evidencial podría haber sido positivamente seleccionada y favorecida porque la actuación partiendo de esta conexión fáctica, que se da realmente, fortalece en general a la adaptabilidad. No estoy sugiriendo que lo que se selecciona positivamente sea la capacidad para reconocer conexiones racionales válidas que existen independientemente, sino que hay una conexión fáctica y que lo que se dio fue una selección entre organismos de la apariencia de validez de este tipo de conexión, que se seleccionó la aprehensión de este tipo de conexión y el que tal aprehensión diera lugar a determinadas creencias, inferencias, etc., ulteriores. Hay selección del reconocimiento de la validez de ciertos tipos de conexión que *son* fácticos, hay, esto es, una presión selectiva para que nos parezcan *más* que fácticos.[4]

Si, con suficiente frecuencia, muestras de un cierto tipo se parecieran a sus poblaciones, entonces la generalización que va de las muestras a la población, o al siguiente miembro con el que nos encontramos, arrojaría frecuentemente verdades; y seres para los que esas inferencias parecieran obvias y evidentes llegarían frecuentemente a esas verdades. Este ejemplo tiene que ver con un proceso general de inferencia inductiva. (Leda Cosmides y John Tooby han investigado la posibilidad de mecanismos inferenciales especializados efectivos para tipos particulares de situaciones frecuentemente recurrentes a lo largo de nuestra historia evolucionaria y, por lo mismo, seleccionados positivamente.)[5] Obsérvese que esta selección evolucionaria podría ser un ejemplo del efecto de Baldwin.[6] En este caso particular, aquellos para quienes «armar» una conexión esté más cerca de ser evidente la aprenderán más rápidamente, ganando así una ventaja selectiva, y dejarán una prole distribuida alrededor del grado en que ellos mismos consideran evidente esa conexión. A lo largo de las generaciones, pues, puede haber un movimiento en el sentido de hallar esa conexión más y más evidente.

Nótese que este punto de vista nos deja —como antes— con el problema de la inducción; una cierta conexión fáctica se dio en el pasado y la selección nos convirtió en organismos que la ven como una

base válida de inferencia, pero ¿continuará manteniéndose esa conexión fáctica ahora y en el futuro? El enunciado de que continuará manteniéndose podría hallarse él mismo en una conexión aparentemente válida con otros hechos que hemos conocido, pero ¿continuará manteniéndose *esta* ulterior conexión (quizás es la conexión con la que empezamos)? Que continúe manteniéndose no queda garantizado por el hecho de que nos parezca evidente a nosotros, pues el que nos lo parezca fue causado por su mantenimiento en el pasado.

Por lo demás, que algo nos parezca evidentemente verdadero no garantiza, en este enfoque, que siempre fue, estrictamente, verdadero. Considérese, por analogía, lo que ahora decimos de la geometría euclídea: es suficientemente verdadera para casi todos los propósitos prácticos; es indetectablemente diferente, en la pequeña escala, de los espacios de curvatura constante pequeña; pero, estrictamente, no es verdadera (para, como decimos, el «espacio físico»). Si hubo selección para que la geometría euclídea pareciera evidentemente verdadera, esto habría sido útil para nuestros ancestros. Nada se habría ganado con la selección de alguna otra geometría. Creer en una geometría alternativa y proceder a inferencias automáticas acordes con *ella* no habría significado ninguna ventaja selectiva (en *este* medio). Pues el que esa geometría alternativa pareciera evidente habría entrañado grandes costes en términos de asignación de recursos neurofisiológicos. E «intuir» esa geometría alternativa no habría estado al alcance de lo que podría haber producido la mutación aleatoria (a falta de presiones selectivas graduales) a partir de la dotación genética que existía entonces. Esta geometría alternativa, literalmente verdadera, no habría sido seleccionada positivamente como lo que nos habría de parecer evidente a nosotros. Dado que la geometría euclídea se aproxima mucho a la verdad, y dadas las ventajas que acompañan al hecho de que nos parezca evidente —ventajas que incluyen la velocidad de inferencia, la creencia de verdades (aproximadas) serviciales y la evitación de otras falsedades más desviadas—, podemos imaginar que la apariencia de evidencia de la geometría euclídea fuera seleccionada positivamente; nos resulta imaginable la selección de esta geometría como nuestra forma de sensibilidad. Ello no obstante, la geometría euclídea es, según creemos ahora, y hablando estrictamente, falsa como teoría del espacio físico. (¿Y sobre qué otra cosa, se ha preguntado Hilary Putnam, se ha supuesto alguna vez que versaba?) No digo que esta historia evolucionaria *sea* la conjetura verdadera acerca de por qué la geometría euclídea parece evidente. Se trata sólo de una analogía para establecer algo que ya sabemos en general, pero que tendemos a olvidar en el ámbito

de las *razones*, la *inferencia* y el *apoyo evidencial*, a saber: que la aparente evidencia de que se dé una conexión (en virtud de algún otro rasgo o relación estructural manifiesta) no es garantía de que se dé de hecho.

¿Diremos algo parecido acerca de la «evidencia» de las reglas deductivas de inferencia y de los mismos principios de la lógica? ¿Son necesarios, o acaso todo el conocimiento *a priori* tradicional ha de pasar al arcón evolucionario? Algunos autores han sostenido que los principios de la lógica no son ni necesarios ni cognoscibles *a priori*; aun si ahora no disponemos de una fórmula alternativa, ciertos fenómenos rebeldes —de la mecánica cuántica, pongamos por caso— podrían llevarnos incluso a la revisión de nuestros principios de lógica.* Yo digo algo más modesto. Para explicar por qué tales principios nos parecen evidentes, no es necesario apelar a su necesidad. Podría bastar con que fueran verdaderos, aun si sólo lo fueran contingentemente, incluso podría bastar con que fueran «suficientemente verdaderos» —recuérdese el ejemplo de la geometría euclídea— y que se hubieran mantenido duraderamente verdaderos el tiempo suficiente como para que hubieran dejado huella en nuestra dotación genética.** Esta posición no deja el flanco abierto a la cogente

* Véase Hilary Putnam, «Three-Valued Logic» y «The Logic of Quantum Mechanics», reproducidos en sus *Philosophical Papers*, vol. 1: *Mathematics, Matter and Method* (Cambridge: Cambridge Univ. Press, 1975), págs. 166-197, y W.V. Quine, *Philosophy of Logic* (Englewood Cliffs, N.J.: Prentice-Hall, 1970) [trad. cast.: *Filosofía de la lógica*, Madrid, Alianza, [5]1991], págs. 85-86, 100. Si la lógica y las matemáticas están en solución de continuidad con, y son una parte de, la ciencia empírica, como sostiene Quine, entonces ¿por qué no buscamos *explicaciones* científicas más profundas de por qué rigen esas leyes de la lógica o esas teorías matemáticas? Los físicos buscan de continuo explicaciones cada vez más profundas de las leyes más profundas conocidas en cada momento, pero los lógicos no llevan a cabo ninguna actividad parecida. Resultaría implausible sostener que esto es a causa de que los lógicos ya han descubierto las leyes más fundamentales. Pues, ¿por qué habría ocurrido esto tan tempranamente en la historia de la lógica y no habría ocurrido aún en física? En una conversación con Quine, me precisó que en *The Roots of Reference* (La Salle, Ill.: Open Court, 1973) [trad. cast.: *Las raíces de la referencia*, Madrid, Alianza, 1988], págs. 76-78, lo que él sostiene es que algunas leyes de la lógica son analíticas; se aprende su verdad según aprendemos el significado de las palabras que las componen. (¡De acuerdo, Quine!) Pero esto no parece suficiente para dar cuenta de por qué no continúa en lógica la búsqueda explicativa de verdades que anden por debajo de las verdades del cálculo proposicional o de la teoría de la cuantificación; la iteración de estos pasos individualmente cortos y conservadores de verdad hasta formar largas cadenas de inferencia conserva también la verdad, ¿y no es esto acaso un hecho matemático (no analítico)?

** Pero la discusión podría continuar: ¿qué explica por qué son verdaderos y se han mantenido verdaderos el tiempo suficiente para tener efectos evolucionarios

objeción de W.V. Quine, según la cual no todas las verdades lógicas pueden deber su verdad a la convención, puesto que necesitamos apelar a los principios de la lógica mismos para derivar consecuencias infinitas de las convenciones.[7] Hemos sugerido que los principios de la lógica se mantienen verdaderos —suficientemente verdaderos, en cualquier caso, y quizá, por todo lo que sabemos, contingentemente— y que el proceso de evolución inculca (no la verdad de los principios de la lógica, sino) su aparente evidencia. De manera que no hay obstáculo para presumir su validez en punto a derivar las consecuencias de que hayan sido inculcados como evidentes. La fuerza y la profundidad de nuestras intuiciones acerca de determinados enunciados no puede usarse como prueba evidencial robusta en favor de su necesidad si esos enunciados son enunciados de un tipo que, de tratarse de hechos contingentes, hubieran podido llevar a la selección positiva de intuiciones muy fuertes sobre su evidencia.[8]

Los filósofos se han enfrentado a la tarea de fundamentar la Razón, de fundamentar lo que tomamos por evidente. El problema humano de la inducción consistía en encontrar un argumento *racional* para concluir que la razón, o la parte de ella incorporada en el razonamiento inductivo, (probablemente) funciona. Aun si pudiera resolverse este problema de Hume,[9] subsistiría la cuestión de cuál sea el fundamento de *este* argumento racional, la cuestión, esto es, de por qué deberíamos confiar en *algún* argumento racional. Éste es el problema al que se enfrentó Descartes —¿por qué las proposiciones que resultan evidentes juzgadas a la luz natural de la razón se corresponden con la realidad?—, y ha dado pie a una extensa literatura sobre el «círculo cartesiano».[10] (Es interesante observar que, en último término, Descartes fundamentó su confianza en el argumento racional en otro ser: Dios.) Kant sostuvo que los racionalistas no podrían probar por qué nuestro conocimiento o nuestra intuición, nuestra «razón» en mi sentido, se conformarían con los objetos, y sugirió —en eso consistió su «revolución copernicana»— que los objetos deben conformarse a nuestro conocimiento, a la constitución de la facultad de nuestra intuición.[11] (De aquí que nuestro conocimiento no lo sea de las cosas en sí, sino sólo de la realidad empírica, pues ésta está modelada por nuestra constitución.)

---

en nuestro sentido de lo que resulta evidente? Si el punto de vista contingente no tiene ninguna hipótesis explicativa más profunda que proponer, el propugnador del punto de vista de la necesidad tampoco puede limitarse a decir que los principios se han mantenido tanto tiempo porque son necesarios, y puede invadirle la perplejidad a la hora de intentar explicar por qué *esos principios* son necesarios.

Si razón y hechos fueran factores independientes, dijo Kant, entonces los racionalistas no podrían producir ninguna razón convincente de la correspondencia entre ambos. ¿Por qué habrían de estar correlacionadas estas dos variables independientes? De manera que propuso que los hechos (empíricos) no fueran una variable independiente; su dependencia respecto de la razón explicaría la correlación y la correspondencia entre ambos. Pero hay una tercera alternativa: que sea la *razón* la variable dependiente, modelada por los hechos, y que su dependencia respecto de los hechos explique la correlación y la correspondencia entre ambos. Nuestra hipótesis evolucionaria presenta precisamente esta alternativa. La razón nos dice cosas sobre la realidad porque la realidad modela la razón, seleccionando lo que a ella ha de resultarle evidente.

Este punto de vista, como queda dicho, puede explicar sólo la correlación pasada; no puede garantizar que los hechos futuros cesarán con la razón presente. Y la misma explicación evolucionaria es algo a lo que llegamos, en parte, usando la razón en apoyo de la teoría evolucionaria en general y, en particular, de su aplicación a este caso. De aquí que no suministre una justificación de la razón independiente de la razón y que, aunque fundamente la razón en hechos independientes de la razón, no aceptemos esta fundamentación de un modo independiente de nuestra razón. Por eso el enfoque adoptado no es parte de la filosofía primera; es parte de nuestro punto de vista científico actual.[12] No se propone satisfacer el criterio kantiano, según el cual «cualquier cosa que guarde alguna semejanza con una hipótesis ha de tratarse como si pasara de contrabando».[13]

Ya va dicho que el mejor modo de entender la filosofía sería como amor a la razón —no como amor a la sabiduría, sino como amor al razonamiento—. Incluso los escépticos griegos y los empiristas británicos la sirven y la honran en sus razonamientos, por mucho que tiendan éstos a disminuir o a socavar la autoridad de la razón y del razonamiento mismo —bordeando así o enfrentándose a una paradoja pragmática—. El intento de fundamentación de la razón por parte del filósofo es su esfuerzo por proteger su amor. (¿O por asegurarse de que su amor le seguirá siendo fiel?) ¿Tendrá la aceptación de esta explicación evolucionaria del poder y de la bella faz de la razón el efecto de amenguar este amor? No está claro que haya de tener ese efecto. ¿Mengua el valor de nuestros ojos y oídos cuando descubrimos que estos órganos perceptivos tienen una explicación evolucionaria? Ello es que algunos filósofos han prestado oídos a la voz de la razón como si ésta anunciara lo necesario contrastándolo con lo contingente, y lo necesario habría de dar acceso a más

que lo que la realidad puede abarcar. Esos filósofos se sentirán privados del especialísimo solaz que les había proporcionado su amor.

El enfoque evolucionario muestra por qué algo podría llegar a resultarnos evidente en su misma apariencia. Dentro del modelo de una retícula de componentes de varios pesos progrealimentadores y sujetos a una regla de corrección de errores, las razones constituyen una categoría más amplia; las conexiones fácticas aprehendidas se reflejan en eslabonamientos y pesos cambiantes. De manera que no estamos reducidos a las razones que la evolución nos ha inculcado. Por fortuna. La evolución podría inculcarnos como evidente algo que es sólo una aproximación a la verdad, y esto podría resultar posteriormente inadecuado para nuestros propósitos. Ocurre también que, dado que no se puede llegar a la exactitud consumada —al menos, a un coste razonable— la evolución ha podido favorecer una disposición a ciertos errores más que a otros. Creer falsamente que no hay un tigre enfrente y mantenerse quieto puede tener efectos más perniciosos en la adaptación que creer falsamente que hay un tigre enfrente y echar a correr innecesariamente. Puede, pues, haber selección evolucionaria de mecanismos que están más prontos a cometer el segundo tipo de error.[14] En diferentes circunstancias, evitar un error dado puede ser menos importante. Un sistema adaptable de procesamiento, con pesos iniciales modificables, será capaz de lograr una mayor exactitud.

Recuérdese nuestra anterior distinción entre (1) que *p* sea lo que hay que creer racionalmente y (2) que creer *p* sea lo que racionalmente hay que hacer. La selección natural opera sobre el segundo miembro de esa distinción, y sólo en la medida en que haya una correlación entre los dos miembros nos proporcionará mecanismos cognitivos engranados con el primero. Más precisamente, la selección natural operará por lo pronto sobre: (3) la acción *A* es la cosa más robustecedora de la adaptación que se puede hacer; o más bien, *A* surge de capacidades cuyo ejercicio, en general, ha mejorado la adaptación inclusiva. Puesto que lo que ustedes hacen será un producto de sus creencias y de sus motivaciones y utilidades, hay un margen en el modo de realización de las predisposiciones a la conducta. De aquí que pudieran evitarse algunos peligros no mediante mecanismos de creencia que se apresuraran a interpretar datos ambiguos sobre la presencia de algo peligroso, sino mediante mecanismos motivacionales, por ejemplo, la repugnancia que lleva a evitar las serpientes.

El robustecimiento de la adaptación inclusiva genera una selección positiva de la verdad aproximada, no de la verdad estricta. Sa-

biendo eso, podemos *perfilar* nuestro objetivo y sus procedimientos. Si se da selección positiva de la servicialidad de una creencia para la acción, y si la verdad es la propiedad que en general anda por debajo de la servicialidad, podemos proponer procedimientos que se centren en la verdad, no sólo en la servicialidad. También podemos perfilar la noción de verdad. Quizá no *toda* servicialidad sea signo de verdad, ni ande ésta siempre por debajo de aquélla —la misma teorización evolucionaria puede decirnos que distintos *tipos* de cosas pueden andar por debajo de la servicialidad—, de modo que podríamos decir que la verdad anda por debajo de una subclase de servicialidad. Cuando nos hacemos conscientes de eso, podemos mejorar la exactitud de nuestros procedimientos.

Consideremos, de nuevo, la conexión entre la fiabilidad con la que un proceso de formación de creencias genera verdades y la adquisición de creencias basada en razones. Si la evolución selecciona mecanismos fiables de formación de creencias, y si creer por razones es un componente de algunos de estos mecanismos fiables, entonces los organismos resultantes pueden preocuparse por y centrarse en las razones más que en la fiabilidad. El centrarse en razones es la vía por la que son conducidos a la fiabilidad, pero ellos no se centrarán en la fiabilidad misma. Análogamente, en el pasado fueron seleccionados mecanismos psicológicos estadísticamente correlacionados con la maximización de la adaptación inclusiva, no una preocupación por la misma adaptación inclusiva. En situaciones en las que la fiabilidad y las razones entran en conflicto, y la persona se da cuenta de ello, la persona puede apoyarse más en las razones que en la fiabilidad; y cuando sólo se da fiabilidad, puede parecer insuficiente. Precisamente: eso es lo que se seleccionó con vistas a la acción. La preocupación por las razones flota ahora libremente a causa de su correlación en el pasado con una vía fiable hacia la verdad.

Adaptación y función

Veamos más de cerca la estructura y los perfiles de la explicación evolucionaria. La evolución, nos enseña la bibliografía por ella generada, tiene que ver con variación heredable en la adaptación, con la transmisión a la progenie de características paternas que varían de organismo a organismo y desempeñan un papel en la reproducción diferencial no aleatoria.[15] La adaptación no consiste en el éxito reproductivo real —pequeños accidentes pueden afectarla—,

y así han podido argumentar Susan Mills y John Beatty que consiste en la propensión (probabilista) de un organismo a sobrevivir y a reproducirse.[16]

Sugiero que entendamos la atribución de una mayor adaptación como un enunciado existencialmente cuantificado. Decir que un organismo *A* está más adaptado que un organismo *B* en el medio *E* es decir que existe(n) algún(os) rasgo(s) fenotípico(s) heredable(s) *F*, tal(es) que *F* explica (por causarlos) el mayor éxito reproductivo de *A* respecto de *B* en *E*. De aquí que decir que *A* tiene mayor éxito reproductivo que *B* porque *A* es más apto que *B* *no* sea una tautología, aunque se diga que *A* tiene mayor éxito reproductivo que *B* en *E* porque existe algún rasgo fenotípico heredable *F* que explica este mayor éxito reproductivo. Podría haber otras explicaciones de este mayor éxito reproductivo, por ejemplo, el azar.

La cuantificación existencial centra la atención en el nivel intermedio de rasgos fenotípicos y en las actividades y funciones necesarias para la supervivencia y la reproducción de una prole viable, actividades y funciones que esos rasgos traen consigo. El rasgo fenotípico *F* opera siempre a través de una de estas funciones generales intermedias *G* (tales como evitar predadores, encontrar comida, transformar energía, atraer pareja, conseguir suficiente humedad, mantener el calor, etc.), permitiendo que esa función o esa actividad se cumplan mejor, más eficientemente, etc. Un listado de *todas* estas funciones intermedias daría más contenido concreto a la teoría de la adaptación; a falta de él, cuantas más funciones podamos enumerar, más determinado será el contenido que daremos a la teoría. Una definición de la adaptación podría incorporar una referencia a este nivel intermedio (intermedio entre el rasgo fenotípico *F* y el éxito reproductivo mismo) *G* de actividad y de función, en el que cada actividad y función *Gi* es tal que, *ceteris paribus*, cuanto mejor cumple *Gi* el organismo, mayor es la probabilidad de su éxito reproductivo. Así, decir que un organismo *A* es más apto que el organismo *B* en el medio *E* es decir que existe algún rasgo fenotípico heredable *F* y alguna función *Gi* de nivel intermedio, tales que *F* causa el robustecido cumplimiento de *Gi* y, por lo mismo, causa (quizá probabilísticamente) el mayor éxito reproductivo de *A* respecto de *B* en *E*.

Hay una complicación adicional a la que debemos prestar atención. (Sin duda, necesitaríamos añadir más detalles para lidiar con complicaciones ulteriores.) Puesto que no es necesario que el organismo más apto tenga un éxito reproductivo realmente mayor (debido al azar de la mortalidad, al azar de la mutación, y aun a otros

factores), quizá deberíamos decir más bien que los rasgos fenotípicos heredables de los organismos más aptos explicarían el mayor éxito reproductivo si se diera ese éxito. En tal caso, el que *A* sea más apto que *B* en *E* podría explicarse del modo que sigue. Hay algún rasgo fenotípico heredable *F* y alguna función intermedia *Gi*, tales que: (a) *A* es reproductivamente más exitoso que *B* en *E*, y *F* causa el robustecido cumplimiento de *Gi*, y así, explica (quizá probabilísticamente) el mayor éxito reproductivo de *A* respecto de *B* en *E*; *o bien* (b) *B* tiene igual o mayor éxito reproductivo que *A* en *E*, y no hay ningún rasgo fenotípico *F'*, tal que *F'* explique el igual o mayor éxito reproductivo de *B* respecto de *A* en *E*, y si *A fuera* reproductivamente más exitoso que *B* en *E*, entonces esto quedaría explicado porque *A'* estaría en posesión del rasgo heredable *F*.

Podría ser mejor, sin embargo, evitar estas complicaciones contrafácticas y esas cláusulas definicionales, pues esas condiciones se prestan a complejos contraejemplos. En cambio, podemos sostener que hay mayor adaptación sólo cuando realmente se da un éxito reproductivo mayor, aunque el éxito reproductivo mayor puede ser causado por algo distinto de la mayor adaptación. Esto sería tanto como decir (aproximadamente) que mayor adaptación es mayor éxito reproductivo causado (probabilísticamente) por rasgos fenotípicos heredables (a través del cumplimiento de algunas funciones intermedias). Sin éxito reproductivo real mayor, no hay mayor adaptación. Sin embargo, los otros organismos que *también* exhiben un mayor éxito reproductivo no cuentan automáticamente como organismos más aptos, porque no necesariamente su mayor éxito reproductivo ha sido (probabilísticamente) causado por algún rasgo fenotípico heredable. (Aun cuando esta construcción convertiría en tautológica la supervivencia del más apto, la aptitud de los supervivientes no lo sería. Se podría añadir entonces la conjetura empírica de que, para la parte preponderante, los supervivientes *son* más aptos.)

La noción de adaptación que hemos descrito es una noción comparativa («más apto que ____ en el medio *E*»). Para varios propósitos, resulta necesaria una medida más fuerte de la adaptación que la mera comparación, pero no es necesario que esta medida más fuerte sea un sencillo número real en vez de alguna entidad matemática más compleja —un *n*-tuplo ordenado, un vector, una matriz, una estructura arbórea de matrices, o cualquier otra—. John Beatty y Susan Finsen aseveran que las probabilidades que tenga un organismo de dejar exactamente *n* descendientes en la próxima generación constituyen por sí mismas una estrategia reproductiva de ese organismo, y es posible que haya una selección entre esas estrategias.[17] Yo

sugiero, en cambio, que consideremos un vector [*p*0, *p*1, *p*2, ..., *pn*], en el que *pi* es la probabilidad de dejar exactamente *n* descendientes en la próxima generación (y esa probabilidad es cero para todos los *i* mayores que *n*). ¿Por qué desechar una parte de esa información?[18] Además del vector para la generación 0, queremos considerar también la adaptación más a largo plazo de un organismo, sus varias probabilidades de dejar exactamente *i* descendientes en la segunda generación, probabilidades que serán una función de su vector previo de una generación junto con los vectores de una generación de cada uno de sus (potenciales) descendientes; y así sucesivamente para las siguientes generaciones. Aquí hay un rompecabezas, empero. Las concreciones particulares de esas estructuras matemáticas serán características biográficas, entre las cuales habrá selección. ¿En términos de *qué* noción de adaptación podremos entonces *explicar* la mayor adaptación de una de esas concreciones respecto de otra en el medio *E*? Si una formación de vectores domina a otra, la cuestión es sencilla; pero a falta de dominación, abundan las complicaciones. El curso real de la historia podría convertirse en un sendero *improbable* a través de la matriz, e incluso la noción de la supervivencia *probable* del más apto (de acuerdo con nuestro primer enfoque) podría carecer de un sentido claro. ¿Evitamos ese rompecabezas debido a que la particular noción de adaptación usada depende de los particulares fenómenos que han de ser explicados?

La evolución que conocemos tiene que ver no sólo con la variación heredable de adaptación, sino también con la imperfecta replicación del material genético en la prole. La mayoría de las mutaciones son perniciosas, y los organismos (de cierta complejidad) contienen mecanismos para compilar y corregir los errores de replicación; pero esos mecanismos son imperfectos. Aun si tal perfección fuera posible, podría haber sido seleccionada negativamente. Todos nosotros somos descendientes de mutaciones que se las arreglaron bien compitiendo con parientes mejor replicados. Sin embargo, si en el gobierno de la replicación operara un mecanismo de corrección de errores *demasiado* defectuoso, no conseguiría preservarse la mutación permitida por ese mecanismo, ni se preservaría *él mismo* —se trata de un ingenio heredado— a lo largo de las generaciones. Por eso el grado de exactitud del mecanismo corrector de errores está él mismo sometido a selección, y el margen permitido por nuestro mecanismo común real no debería considerarse un defecto —los organismos con mecanismos de corrección de errores más perfectos aún son protozoarios—.*

* Podría resultar útil comparar el mecanismo genético de corrección de errores con una máquina de Turing dotada de un *scanner* que se moviera casilla a casilla,

Creo también que hay espacio para un concepto de adaptación más general que el que se limita a comparar el éxito diferencial de los organismos reproductivos. Consideremos la cuestión de por qué hay entidades reproductoras (u organismos). Una vez que existen los organismos reproductores, proliferan. ¿No debería haber una noción más general de adaptación que abarcara las ventajas que puedan tener las cosas vivas reproductoras en comparación con las cosas no vivas? Quizá la medida que andamos buscando debería hablar de la competición entre átomos y moléculas: ¿tienen éxito los organismos vivos en punto a incorporarlos a las cosas vivas? Podría, así, pues, llegarse a una *ratio* entre la biomasa y la no biomasa en el medio material local (cerrado). (Me ha contado Richard Lewontin que en botánica no siempre está claro o es relevante el *número* de organismos, de manera que también allí podría ser útil una noción de adaptación que se centrara en otras medidas materiales.) ¿Hay un límite teórico para la cantidad de materia del universo que puede incorporarse a la biomasa? ¿Localmente, en la tierra, estamos aún en la parte ascendente de la curva? ¿Qué forma ha cobrado esa curva en el transcurso del tiempo? Nuevas e interesantes cuestiones pueden aparecer cuando formulamos un concepto más general que ilumina por qué la vida reproductiva sobrevive y prolifera, un concepto que incluye como una concreción particular la noción biológica usual de adaptación.

Hemos considerado a la racionalidad como una adaptación biológica con una función. ¿Qué es una función? ¿Cómo se inserta la noción de función en un marco biológico y evolucionario? Ernest Nagel ofrece un iluminador análisis de un sistema homeostático como es el sistema regulador de la temperatura de nuestros cuerpos. Tal sistema mantiene el valor de una de las variables de estado *V* dentro de un cierto espectro en un cierto medio, de manera que cuando se causa la desviación de *V* a cierta distancia (pero no a cualquier distancia arbitrariamente grande) más allá de ese espectro, los valores de las otras variables compensan eso, modificándose para

---

que procediera a los cambios ineludibles, añadiendo en caso necesario nuevas casillas a la cinta procedentes del material circundante. ¿Podrían alguna estructura formal y los resultados de la teoría de la computación iluminar esta porción de la biología? Quizá necesitemos una teoría de una máquina de Turing, ligeramente imperfecta, que a veces computara mal. Obsérvese la analogía autorreferencial con el fenómeno de un mecanismo de corrección de errores que reparara, entre otros materiales, el que fuera causalmente responsable de la (siguiente) reproducción de *este* mecanismo de corrección de errores.

hacer regresar a *V* de nuevo hasta el interior del espectro en cuestión.[19] Nagel presenta esto como un análisis de un sistema teleológico u orientado a un objetivo, siendo el objetivo o la función del sistema mantener la variable *V* dentro del espectro. De acuerdo con este enfoque, cualquier otra variable *V'* universalmente asociada a *V* constituiría también el estado-objetivo de ese sistema homeostático, una consecuencia contraintuitiva que podría evitarse atendiendo a la *explicación* de por qué el sistema mantiene *V* dentro de ese espectro. No cualquier cosa con una función, empero, es un sistema homeostático. La función de una silla de comedor podría ser la de aguantar a una persona sentada a la mesa, pero no por cumplir mal su función modificará los valores de algunas de sus variables de estado para poder constituir un mejor soporte corporal.

Para lidiar con este y otros casos, Larry Wright propone que la función de algo ayude a explicar por qué existe ese algo: la función de *X* es *Z* cuando *Z* es una consecuencia o resultado de la existencia de *X*, y *X* existe porque existe *Z*.[20] Christopher Boorse objeta que si un científico construye una manguera con una fisura y muere por el consiguiente escape de gas antes de que pueda reparar la fisura, ello no nos permite decir que sea una función de la fisura el dejar escapar el gas aunque de la fisura resulta el escape y continúa existiendo gracias al escape.[21]

Permítasenos una mirada nueva al asunto. *Z* es una función de *X* cuando *Z* es un efecto de *X* y *X* estaba diseñado o moldeado (o mantenido) para tener este efecto *Z*. Tal diseño o moldeamiento puede resultar tanto de un diseñador humano, como de un proceso evolucionario. En ambos casos, el diseño mismo parece ser un proceso homeostático cuyo objetivo es que *X* produzca *Z*. Un diseñador humano moldea una silla para que aguante a una persona, modificando rasgos de esa silla —en la planificación o en la fabricación— para que su función pueda cumplirse efectivamente. A lo largo de generaciones, la evolución moldea organismos y órganos corporales para conseguir ciertos efectos más efectivamente; selecciona positivamente organismos para los cuales vale eso. (Ni que decir tiene que hay otros procesos, además de la selección adaptativa, que operan en la evolución, por ejemplo, la deriva genética. Y entender la evolución, en una parte significativa, como un mecanismo homeostático no implica considerarla como un mecanismo de optimización.) Podemos combinar los puntos de vista de Nagel y Wright —tal es la sugerencia— para dar una imagen más completa de la función. *Z* es una función de *X* cuando *Z* es una consecuencia (efecto, resultado, propiedad) de *X*, *y* la producción misma de *Z* por *X* es el estado-objetivo

de algún mecanismo homeostático *M* que satisface los criterios del análisis de Nagel, y *X* fue producida o es mantenida por este mecanismo homeostático *M* (a través de su búsqueda del objetivo: la producción de *Z* por *X*). (Se podría eliminar la primera cláusula que exige que *X* tenga realmente el efecto *Z*: no todas las funciones pueden cumplirse.)

Este enfoque explica por qué los biólogos no dicen que la función del ADN basura es no hacer nada o ser más caro de eliminar que de mantener, o que la función de un perturbador de segregación es romper la meiosis.[22] Aunque estos *son* efectos, el ADN basura y los perturbadores de la segregación no fueron modelados por un proceso homeostático para tener estos efectos. Ningún proceso homeostático tendió a esos efectos o los seleccionó positivamente. Obsérvese que el punto de vista propuesto aquí no convierte automáticamente al estado-objetivo de un sistema homeostático en su función. Un termostato creado por una combinación accidental de elementos no tendría la *función* de regular la temperatura aun si éste fuera su efecto. En mi enfoque, el mecanismo homeostático es el diseñador, no el objeto diseñado, y *su* objetivo es que alguna otra cosa *X* produzca el efecto *Z*, alguna otra cosa *X* que él crea o mantiene, y es a *X* a quien se atribuye la función. ¿Qué efectos de *X* son su función? Aquellos para los que fue producida o diseñada por (o es mantenida a causa de) algún mecanismo homeostático. (Puesto que algunos efectos laterales serán coextensivos con los efectos que fueron seleccionados positivamente, o bien una atribución de función será vinculada a una *explicación* de por qué un efecto fue seleccionado positivamente —en donde «explicación» no es una noción extensional—, o bien el enfoque aquí presentado será una condición necesaria pero no suficiente para que un efecto sea una función.)

Los procesos de evolución producen entidades con funciones, pero la evolución misma tendría una función —por ejemplo, la de producir entidades con funciones— sólo si hubiera algunos *otros* procesos homeostáticos que produjeran o mantuvieran la evolución para tener este efecto. Nótese que no es una consecuencia del presente punto de vista el que las cosas (moralmente) *deberían* cumplir sus funciones; seres creados y moldeados por un mecanismo homeostático para ser esclavos y trabajar duro tendrían (en este enfoque) esta función, pero sin embargo deberían rebelarse. Obsérvese también que alguna cosa existente que no fue moldeada por un proceso homeostático podría en algún momento empezar a ser *usada* para un propósito —por ejemplo, una gran piedra plana como mesa de picnic—. En tal caso podríamos decir que se le ha *dado* una función;

hacemos lo mismo con ella que con algo que haya adquirido una función por la vía de un mecanismo homeostático. *Funciona como* una mesa, aunque no sea ésa su función. Pero si ahora moldeamos o mantenemos el objeto para retenerlo dentro del espectro de características que lo hacen utilizable para este efecto, limpiándole el musgo y quitándole la hojarasca, por ejemplo, entonces su continuada condición dirigida a tener este efecto —ser usado como una mesa de picnic— *es* el resultado de un mecanismo homeostático, consiguiendo así ahora su función.

LA FUNCIÓN DE LA RACIONALIDAD

¿Son las razones mismas evidencia* para aquello para lo que son razones? ¿Cuál es *su* función? La respuesta puede parecer evidente. Las razones están conectadas con la verdad de aquello para lo que son razones —tal es la conexión fáctica—, de manera que creer por razones es una vía hacia la creencia de la verdad. De acuerdo con nuestro análisis de la función, la función de «creer o actuar por razones» será algún rasgo o efecto que esto tenga, rasgo o efecto que algún mecanismo homeostático subyacente «se propone» que tenga. Puesto que la racionalidad toma en cuenta (y actúa de acuerdo con) razones, lo que sea la racionalidad, la función que tenga, dependerá de un hecho acerca del mundo, a saber: qué mecanismo(s) homeostático(s) opera(ron), y en pos de qué objetivo, al moldearnos para actuar y creer sobre la base de razones.

El primer mecanismo homeostático que hay que considerar es el proceso evolucionario que opera a través de la selección natural. ¿Fue el creer o el actuar por razones un rasgo seleccionado positivamente, y si es así, por qué? Obsérvese que podría haber habido selección positiva de rasgos que arrojaran el creer o el actuar por razones como un *producto lateral,* sin seleccionarlos positivamente de un modo directo. En tal caso, creer o actuar por razones no tendría una función evolucionaria; podría no haber ninguna propiedad *P,* tal que un mecanismo homeostático evolucionario tuviera como su estado-objetivo (mantener) el hecho de que creer o actuar por una razón tiene esta propiedad *P.* Porque el mundo cambia de manera no regular, según dijimos, necesitan los organismos mecanismos adaptativos para responder a circunstancias locales; no pueden car-

* Lo que consiste, según he sugerido, en una doble conexión: una conexión fáctica que también es estructural y parece evidente.

gar con toda esta tarea estructuras permanentes y respuestas preprogramadas, unidas al condicionamiento operante. Las razones y el razonamiento resultarían útiles para un organismo enfrentado a situaciones nuevas que tratara de evitar futuras dificultades. Tal capacidad para la racionalidad, ya haya sido seleccionada directamente o esté a caballo de otras capacidades, muy bien podría servir a un organismo en sus tareas vitales e incrementar su adaptación inclusiva. Esto conferiría a la racionalidad explícita la tarea de lidiar con los hechos y las necesidades cambiantes, y quizá de modificar nuestra conducta filogenética cuando nos apercibiéramos de cambios presentes que la tornaran inadecuada.

La evolución puede habernos inculcado información filogenética acerca de, y pautas de conducta adecuadas a, hechos estables de nuestro pasado evolucionario. No tenemos que pensar explícitamente en (ni siquiera conocer) la alternancia regular del día y la noche; nuestros ritmos corporales cumplen esta tarea por nosotros, como descubren los viajeros con *jet-lag*. Algunos hechos son lo bastante estables para que los organismos puedan prescindir de su conocimiento; la evolución se los habrá grabado. La racionalidad puede tener la función evolucionaria de capacitar a los organismos para lidiar mejor con situaciones presentes nuevas y cambiantes, o con situaciones futuras presagiadas por algunos indicios presentes, posiblemente complejos. Que la racionalidad *pueda* hacer eso es uno de los hechos estables, no un hecho que tengamos que conocer explícitamente. Todo lo que necesitamos es tenerlo grabado en nosotros, como lo está el hecho de la gravedad. La evolución emplea y construye mecanismos alrededor de rasgos ambientales constantes y estables. No es sólo que la gravedad sea una restricción para algunas características —por ejemplo, el tamaño—, sino que la fuerza gravitatoria se utiliza para el funcionamiento de algunos procesos. La fisiología de los astronautas se ve seriamente afectada por la ingravidez prolongada, porque sus mecanismos fisiológicos evolucionaron en un medio de gravedad constante y fueron diseñados para utilizarla y funcionar de consuno con ella. Tales procesos no duplican lo que la gravedad ya hace (y el intento de duplicarlo, en presencia de gravedad, no haría sino producir una fuerza excesiva y resultados dañinos).

Podemos ensayar la siguiente hipótesis. Todos los problemas filosóficos con los que los pensadores filosóficos han batallado largo tiempo, sin éxito evidente —los problemas de la inducción, de las otras mentes, de la existencia del mundo externo, de la justificación de la racionalidad— apuntan a presupuestos que la evolución ha instalado en nosotros. También hay otros problemas, menos familiares.

Mas ¿por qué pensar que nos es necesario resolver esos problemas? Si todos los seres humanos han nacido hasta ahora en medios poblados por otras personas, no hay ninguna necesidad de que aprendan o infieran por sí mismos que se trata de personas con mentes similares a las suyas. Cualquier humano tendría que saber que esto funciona efectivamente; aquellos primos de nuestros ancestros que no consiguieron aprenderlo no dejaron tras de sí descendientes similarmente incipientes. Quienes aprendieron esto más rápidamente tuvieron alguna ventaja, y así, por la vía del efecto Baldwin, la evolución iría inculcando más y más capacidad de aprenderlo hasta que, finalmente, ese conocimiento fue instalado. Análogamente, a aquellos primos de nuestros ancestros que no consiguieron aprender que había un «mundo externo» que existía independientemente, un mundo cuyos objetos seguían sus propias trayectorias o permanecían en su sitio cuando no eran observados, no les fue tan bien como a aquellos que reconocieron rápidamente la obstinación de la realidad. Quienes no consiguieron aprender a generalizar a partir de la experiencia pasada (de una forma adecuada para ellos) sucumbieron a peligros que les dejaron con menos descendencia. Nunca fue función de la racionalidad la de *justificar* estos supuestos que encarnaban estabilidades de nuestro pasado evolucionario, sino utilizar esos supuestos para lidiar con las cambiantes condiciones y los cambiantes problemas *dentro* del marco estable fijado por ellos. No debería sorprendernos que nuestros instrumentos racionales no consigan suministrar razones o «justificación» para esos supuestos. No fueron diseñados para este propósito, ni para el propósito de suministrar razones concluyentes para su *propio* uso.[23]

«La probabilidad es la guía de la vida», dijo el obispo Butler, pero no somos capaces de establecer la racionalidad de actuar, en situaciones particulares, sobre la base de lo que es más probable. En realidad, hasta hace poco, en la mayoría de las teorías corrientes de la probabilidad, no tenía sentido imputar una probabilidad a un acontecimiento particular, y se hubieran necesitado razones para mostrar, por ejemplo, por qué lo que habría de ocurrir en una hipotética secuencia infinita de acontecimientos debería servirnos de guía en un caso particular muy finito. Recientemente, se han formulado interpretaciones de la probabilidad en términos de propensiones, interpretaciones que imputan probabilidades a acontecimientos particulares. Aun así, en esas teorías, y también en la construcción de la probabilidad como un término *teórico* en la ciencia para dar cuenta de determinados fenómenos, nos quedamos sin saber por qué deberíamos actuar la siguiente vez de acuerdo con lo más probable.

La racionalidad de actuar basándose en la probabilidad se expresa en la teoría de la utilidad mediante la condición de Von Neumann-Morgenstern, según la cual si una persona prefiere *x* a *y*, entonces la persona prefiere, de dos mezclas de probabilidades que conceden (sólo) a *x* y a *y* distintas probabilidades, la mezcla que conceda mayor probabilidad a *x*. Si *x* es preferido a *y*, entonces [*px*, (1-*p*)*y* es preferido a *qx*, (1-*q*)*y* si y sólo si *p* es mayor que *q*]. Esta última sentencia tiene exactamente la misma forma que una sentencia carnapiana de reducción,[24] sugiriendo así el proyecto de definir implícitamente la probabilidad en sus términos. En vez de preocuparse por justificar por qué habríamos de actuar basándonos en probabilidades, defínanse las probabilidades en términos de cómo deberíamos actuar. Este proyecto fue llevado a cabo por L.J. Savage, quien fijó un conjunto de condiciones normativas (y estructurales) para la conducta, para las preferencias entre acciones, suficientes para *definir* una noción de probabilidad personal.[25] Por ejemplo, si una persona que prefiere *x* a *y* elige la acción *A* antes que la *B*, llevando *A* a *x* si ocurre el estado *S*, y a *y* si ocurre el estado *T*, mientras que, al revés, *B* lleva a *y* si ocurre el estado *S* y a *x* si ocurre el estado *T*, entonces podemos entender su elección de *A* como reveladora de que esa persona cree (como *definidora* de su creencia de que) *S* es más probable que *T*. Pero definir la probabilidad en términos de esta *única* preferencia entre acciones puede resultar engorroso. Supongamos que la misma persona también prefiere *z* a *w* y elige la acción *C* antes que la *D*, cuando *C* da lugar a *z* si ocurre el estado *T*, y a *w* si ocurre *S*, esto revelaría que la persona cree más probable *T* que *S*. Pero su anterior preferencia de *A* sobre *B* revelaba que cree más probable *S* que *T*. Si se trata de definir las probabilidades personales en términos de las preferencias entre acciones, hay que evitar ese conflicto. De aquí la condición impuesta por Savage: si la persona prefiere *A* a *B* para los particulares resultados *x* e *y*, en donde *x* es preferido a *y*, entonces para *cada z* y para *cada w*, tales que la persona prefiere *z* a *w*, preferirá también *D* a *C*. Sin embargo, no hay ninguna razón independiente para imponer este requisito normativo, a no ser que esa razón reconociera alguna noción independiente de probabilidad distinta de la noción de probabilidad personal y considerara las elecciones de la persona como reveladoras de que actúa fundándose en una probabilidad que cree mayor.[26] (También podría, alternativamente, actuar fundándose en algún principio que no tuviera que ver con probabilidades, o que tuviera que ver con ellas de un modo diferente.) Así, pues, la condición supone que la persona debería tener siempre algunas creencias probabilitarias clara-

mente definidas y debería actuar siempre fundándose en ellas (del modo determinado por el axioma de Von Neumann-Morgenstern). Mas ésta es la auténtica cuestión que anda en juego —por qué es racional actuar basándose en lo más probable—, y esa cuestión no queda eludida por el intento de Savage de definir la probabilidad en términos de acción.[27] El caso es que los recursos de la racionalidad no han suministrado hasta ahora una respuesta satisfactoria para este asunto —por qué, en un caso particular, deberíamos actuar basándonos en lo más probable o creer lo más probable—, por otra parte tan central para la noción de racionalidad instrumental y de racionalidad de la creencia.[28]

Consideremos el intento de Kant para convertir a la conducta regida por principios en el único criterio último de conducta, independientemente de cualesquiera deseos particulares que un individuo pudiera tener. Los principios, empero, son mecanismos —no digo que sean una adaptación evolucionaria— constituidos para funcionar de consuno con deseos ya existentes, algunos de los cuales han sido inculcados biológicamente. De modo que los principios son mecanismos parciales. Pero Kant trunca el contexto en el que los principios pueden funcionar eliminando los cofactores, para funcionar con los cuales están diseñados esos principios; los principios solos deben cargar con toda la tarea, la idea misma de principio. Nuestra anterior discusión de los principios no les confería ese alcance, esa tarea tan desencarnada. No sólo nuestra racionalidad y nuestros principios son parciales, diseñados como están para funcionar de consuno con cosas externas; nosotros, los seres humanos, somos criaturas parciales, no enteramente autónomas. Somos parte del mundo natural, estamos diseñados para trabajar en equipo con otras partes y otros hechos, dependemos de ellos. La memoria humana usa información almacenada en objetos (dispuestos) en el mundo exterior;[29] también somos criaturas físicas que ocupan nichos ecológicos. El enfoque evolucionario de la racionalidad y de sus limitaciones liga con un tema de los escritos de Ludwig Wittgenstein, John Dewey, Martin Heidegger y Michael Polanyi, quienes, por distintas razones, entendieron también que la racionalidad estaba incrustada en un contexto y desempeñaba un papel como uno entre otros componentes, en vez de como un podio externo y autosuficiente desde el que juzgarlo todo.[30]

La explicación evolucionaria de por qué no podemos justificar racionalmente ciertos supuestos —nuestra racionalidad fue diseñada para trabajar en equipo con aquellos hechos, cumpliendo otras funciones— no significa por sí misma una justificación de esos su-

puestos. Así como a la geometría euclídea le bastaba con ser «suficientemente verdadera», así también podría haber quedado fijada la creencia en otras mentes y en un mundo externo independiente (a través del efecto Baldwin), *sin* necesidad de que fuera, estrictamente hablando, verdadera —todo lo que necesitan ser esas creencias es «suficientemente verdaderas»—. De manera que quizá no podamos estar completamente seguros de con qué verdades estamos trabajando en equipo.

Además, que se haya dado una regularidad en el pasado, según nos enseñó Hume, no garantiza que se siga dando (probablemente) en el futuro. Que los hechos pasados llevaran a que fueran instalados en nosotros determinados supuestos que casaban con ellos no significa que aquellos hechos sigan estando vigentes y que aquellos supuestos sigan sirviéndonos. Y aun si *alguna* regularidad que se dio en el pasado continuara dándose, Nelson Goodman observa que hay muchas de esas regularidades que casaron con el pasado y que sin embargo divergirán en el futuro. ¿Cuál de ellas continuará?[31] La evolución sólo selecciona rasgos que han sido útiles hasta ahora. Rasgos que han sido igualmente útiles hasta ahora resultan igualmente favorecidos, por mucho que pueda diferir en el futuro la utilidad de esos rasgos. La evolución no favorece especialmente a la utilidad en el futuro, aunque retrospectivamente pueda decirse que llegó a favorecer lo que acabó siendo útil en este tiempo ahora pasado que previamente fue futuro. Desde esa perspectiva, cuando se dice que la evolución encapsula en nuestra herencia filogenética las regularidades estables del pasado, la cuestión no es simplemente si las regularidades del pasado continuarán dándose, de manera que nuestra herencia seguirá sirviéndonos útilmente, sino si la evolución ha escogido la regularidad «correcta» o nos ha dado «verde» (*green*) en un mundo «pardo» (*grue*). (¿Pueden las reglas de proyección de Goodman interpretarse como criterios de adaptación comparativa?)

Con todo, el hecho de que la racionalidad no fuera diseñada para justificarse a sí misma o para justificar su marco de supuestos no implica que sea *imposible* que lo haga. (Pero el hecho de que los filósofos hayan fracasado hasta ahora da peso a la opinión de que no habría sido eficiente dejar esa justificación, o incluso esa inferencia, al albur de cada persona individualmente considerada.) Tampoco significa que no podamos hallar razones *contrarias* a la irrestricta verdad del marco de supuestos, aun habiendo sido estos últimos evolucionariamente inculcados. Recuérdese el ejemplo de la geometría euclídea, cuya inexactitud pudo descubrirse a pesar de haber sido seleccionada positivamente como «evidente». Para decirlo todo,

hemos podido descubrir eso aplicando otras relaciones de la razón que fueron evolucionariamente inculcadas en nosotros, o más exactamente, modificando esas relaciones de la razón (y fuimos llevados a esas modificaciones sobre la base de otras relaciones ulteriores de la razón). No fue necesario que empezáramos en algún momento con relaciones de la razón que fueran espejos exactos de los hechos acontecidos. Un grupo de relaciones de la razón *aproximadamente* exactas puede modelarse a sí mismo hasta transformarse en un grupo más exacto. Una herramienta aproximadamente precisa puede hacer lo mismo con otra; esta segunda herramienta mejorada puede hacer lo mismo por una tercera; y esta tercera, entonces, puede examinar la primera herramienta y detectar y corregir sus errores. Todo es susceptible de mejora hasta llegar a ser más exacto o preciso de lo que era inicialmente. Con el tiempo, el carácter de cada herramienta inicial podría haber cambiado significativamente, y las tres herramientas juntas podrían ponerse al servicio de la obtención de una cuarta herramienta.[32]

Hay un segundo mecanismo homeostático de moldeamiento a considerar: los procesos merced a los cuales modelan las sociedades a sus miembros. Las capacidades que subyacen a la creencia o a la acción por razones pueden haber sido objeto de selección natural (por cualquier razón); pero, una vez que esas capacidades cobran existencia, la *sociedad* podría haberse apropiado de la oportunidad de producir (de algún modo) miembros racionales. Cuando los científicos sociales hablan de elección racional, su propósito habitual es explicar rasgos de instituciones sociales y mostrar el modo en que los individuos racionales construyen y mantienen la sociedad.[33] Este tipo de trabajo es iluminador, pero también necesitamos investigar cómo y por qué la sociedad fabrica y mantiene miembros racionales. La gente no nace racional. Cualquiera que sea la medida en que algunos procesos racionales *son* un producto de pautas de desarrollo innatamente controladas, esos procesos son moldeados por y descansan en procesos, normas y procedimientos socialmente inculcados. ¿Qué procesos sociales efectúan este moldeamiento y por qué existen? (¿Es acaso sólo por los efectos *causales* que trae consigo el actuar o el creer por razones que la sociedad inculca esos rasgos?)

Las instituciones sociales y las estructuras sociales son mantenidas por las acciones y las elecciones de la gente dadas las restricciones y las estructuras de incentivos con que ésta se enfrenta, y esas restricciones y estructuras vienen fijadas por otras instituciones y por la conducta de la gente que mora en ellas. Como en un inmenso

rompecabezas en el que cada pieza encajara solamente en el espacio que le dejan todas las demás piezas, las acciones de cada persona se realizan en el espacio de restricciones e incentivos que le dejan otras acciones de todos los demás. Las instituciones gozarán de continuidad cuando sean capaces de *reproducirse* a sí mismas, cuando, de consuno con el resto de instituciones de la sociedad, sean capaces de reclutar, entrenar y suministrar incentivos a los miembros y a los funcionarios nuevos. Parte de este entrenamiento puede tener que ver con normas y hábitos de racionalidad, modos de elección y modos de formación de creencias.[34] La institución crearía individuos hasta cierto punto racionales —que serían sensibles a ciertos tipos de incentivos, que percibirían y tomarían en cuenta ciertos tipos de restricciones— para reproducirse a sí mismas como instituciones, quizá ligeramente alteradas, en el siguiente período de tiempo. No es necesario que concibamos a las instituciones como si se trataran de reproducirse a sí mismas. El proceso es un proceso selectivo; aquellas instituciones que (aun trabajando en equipo con otras) no se propagan, reproducen o autorreplican no sobreviven. Ésa es la razón por la que hallamos que casi todas las instituciones existentes emplean medios para reclutar y entrenar a personas nuevas para hacerse cargo de las funciones necesarias para la pervivencia de las instituciones.*

* Además de las funciones sociales de la racionalidad, está también el tópico de su carácter social, de los modos en que la racionalidad es interpersonal. Jürgen Habermas ha sostenido que la racionalidad es imposible sin una abierta disposición a atender a las opiniones procedentes de todas las fuentes y sin la libertad y los recursos que esas potenciales fuentes de opinión necesitarían para poder participar en una discusión razonada. Pero la racionalidad no requiere el más amplio basamento evidencial, computacional, etc. Ese proceso mismo tiene sus costes, de lo que habría de resultar alguna decisión (aproximada) respecto de la cantidad de tiempo y de energía que hay que invertir en cualquier decisión particular o en cualquier formación de creencia. Obsérvese que esos costes son costes personales, no costes peculiarmente epistémicos. Si tales límites no convierten necesariamente en irracional las creencias y decisiones particulares de los individuos —en realidad lo irracional para ellos sería no imponer esos límites—, entonces no resulta claro por qué las limitaciones sociales, la falta de participación democrática en la formación de la opinión, etc., impediría que las concepciones de los individuos en la vida social (o las de la vida social misma, si se dispusiera de una noción de racionalidad social) fueran racionales. Se podría sostener que lo que marca la diferencia es la naturaleza del proceso por el que se decide acerca de esas limitaciones de costes. Sin embargo, en analogía con el individuo que decide por sí mismo cuánto debe investigar, las autoridades de una sociedad no democrática podrían decidir cuán difundidas habrían de ser sus propias fuentes de información y de opinión contraria; así que no resulta claro por qué sus creencias y decisiones no podrían ser racionales. Es verdad que esas autorizaciones deberían considerar también los *posibles* ses-

De aquí que una función significativa de la racionalidad sea la de propagar instituciones en el tiempo hacia etapas institucionales ulteriores, no servir a los intereses de los individuos entrenados y moldeados para ser racionales. (La precisa naturaleza de la institución afectará al posible robustecimiento o debilitamiento, por parte de esa modelación, de la conexión entre razones y fiabilidad en punto a llegar a la verdad.) En *El gen egoísta,* Richard Dawkins consideró a los organismos y a su conducta como ingenios diseñados para servir a la reproducción de los genes.[35] («Una gallina es el camino seguido por un huevo para fabricar otro huevo.») Las presentes reflexiones plantean otra posibilidad; que la racionalidad es moldeada, seleccionada y mantenida no para servir a un nivel por debajo del de los organismos, sino para servir a un nivel *por encima* de ellos: el nivel de las instituciones.* Eso no significa que podamos ignorar el juego de interacción que se da entre los individuos y las instituciones; ambos se moldean mutuamente de un modo que tendrá consecuencias para el moldeamiento futuro de cada uno de ellos.

Habrá a veces hipótesis competitivas sobre el nivel al que ha operado la selección. Considérese el supuesto, corrientemente realizado por los economistas, de la maximización de la riqueza. Este supuesto, más específico que el de la maximización de la utilidad, proporciona contenido detallado a sus teorías. Se apela a la maximización básicamente por su docilidad y poder matemáticos, de manera que podemos considerar el supuesto más débil y más plausible de que la gente se preocupa seriamente de la riqueza (aun si no la maximiza, o aun incluso si no la sitúa lexicográficamente en pri-

---

gos en las limitadas fuentes de información por ellas empleadas, pero también esto podrían remediarlo con un *muestreo* limitado de un conjunto mucho más amplio de opiniones. De modo que no parece que la discusión pública y democrática y la formación de opinión en una sociedad sean condiciones necesarias para la racionalidad de las decisiones y creencias de sus miembros, o de ella misma. Las creencias racionales de los antiguos griegos no se vieron estorbadas por el hecho de que su sociedad incluyera esclavos. Los derechos democráticos tienen una base distinta.

* Ha habido mucha discusión entre biólogos y filósofos de la biología acerca de la unidad de selección o del nivel al que opera la selección. Una de las chispas que encendió esta discusión fue la cuestión de la selección de grupo. Un fenómeno frecuentemente citado en apoyo de la selección de grupo o interdémica, un fenómeno aparentemente inexplicable por la vía de la selección individual, es la limitación de la virulencia del virus mixoma entre los conejos. Vale la pena observar que este fenómeno caería bajo la selección *individual* si hubiera una acción modificadora por parte de los genes para cepas menos virulentas, con el efecto fenotípico de reducir la distorsión de la segregación.

mer lugar). Podría darse a eso una explicación social en términos de las instituciones que moldean las preocupaciones y las motivaciones psicológicas de la gente y del modo en que esas particulares motivaciones contribuyen al funcionamiento y propagación de esas instituciones. Pero hay otra posibilidad. Parece que un fenómeno ampliamente difundido entre las sociedades (pero que se da en los últimos 150 años en las sociedades industrializadas occidentales) es que los ricos tienden a tener más hijos. En las sociedades poligámicas, los ricos (normalmente varones) tienden a tener un mayor número de parejas (mujeres), y en todas las sociedades los individuos ricos pueden haber sido más capaces de defender a su prole de las vicisitudes materiales.[36] También hay evidencia de que en las sociedades de cazadores y recolectores en el siglo XX —aparentemente, el equivalente contemporáneo más cercano a una parte significativa de nuestro pasado evolucionario— los jefes y los individuos ricos tienden a dejar más descendientes. Supongamos que, *ceteris paribus*, la gente con un fuerte deseo de riqueza tiende a acumular más que quienes tienen un deseo menos fuerte de ella; es decir, que es más probable que lo hagan. Si hubiera habido una predisposición psicológica heredable y genéticamente basada a estar (más) preocupado por la riqueza —no digo que esto sea más que una posibilidad—, entonces los individuos que estuvieran en posesión de esa predisposición habrían tendido a generar más descendencia que llegara a la edad reproductiva, y también esa descendencia habría tendido a estar en posesión de la predisposición heredable, y por lo tanto, a acumular más riqueza, y por lo tanto, a tener más hijos. Una predisposición heredable a desear y aspirar a la riqueza habría sido seleccionada positivamente. El porcentaje de quienes estuvieran en posesión de esa predisposición heredable se habría ido incrementando en el curso de las generaciones. (Puesto que no más del 50 por ciento de los individuos puede ser más rico que el promedio, ¿cómo habría de afectar esto al resultante equilibrio del porcentaje de maximizadores de la riqueza en la población si lo combináramos con otros varios factores, tales como el grado de movilidad de la sociedad?) De manera que podría obtenerse así una explicación evolucionaria del supuesto si no de la maximización que hacen los economistas, sí al menos del fuerte deseo de riqueza. ¿Somos tan pocos los que nos preocupamos por las cosas elevadas de la vida porque aquellos contemporáneos de nuestros ancestros que sí lo hicieron dejaron menos descendencia y *nosotros* descendemos de quienes tendieron, en cambio, a preocuparse de las posesiones materiales? Evidentemente, las hipótesis biológicas y las hipótesis institucionales no son excluyentes;

ambos factores pueden interactuar en la producción del fenómeno.*

La racionalidad, ya lo dije antes, es autoconsciente en el sentido de que trata de corregir los sesgos de la información que recibe y los de sus propios procedimientos de razonamiento. Nos preguntábamos antes si la función de la racionalidad no es una función limitada, ubicada como está en un marco de supuestos evolucionariamente inculcados, algunos de los cuales apuntan a problemas filosóficos que somos incapaces de resolver. ¿Deberíamos entender también que librarse de sesgos intelectuales tiene una función delimitada, una función que se cumple dentro del marco de los sesgos fundamentales de una sociedad, según presente y pondere ésta las razones y la información? Me parece ésta una idea engorrosa. ¿Qué alcance tiene nuestra corrección de sesgos? ¿Deberíamos conformarnos con la gama de opciones relativas a creencias y acciones que nos presenta nuestra sociedad y con la gama de valores que ha inculcado con objeto de desbaratar ulteriormente esas opciones, y deberíamos limitarnos entonces a emplear razones con el solo objeto de elegir entre las opciones *dentro* de esa gama? ¿O deberíamos acaso empezar con todas las razones a favor y en contra —entendidas éstas del modo más amplio posible—, corrigiendo cualesquiera sesgos, procedentes de la transmisión social de información y de la ponderación y evaluación social de las razones, que pudiéramos detectar, para luego fundar del mejor modo posible nuestra decisión en esa base no sesgada? (Si este proceso lleva a un empate entre las opcio-

* Los factores biológicos pueden interactuar también con otros tipos de factores sociales tales como la intención paterna de moldear la psicología de sus hijos —sin duda, un fenómeno muy difundido—. Quizá los padres lo hacen por el bienestar de sus hijos, o por propia conveniencia en la interacción, o porque instituciones más amplias les han moldeado para que lo hagan. Pero hay otra posibilidad digna de mención, a saber: que los padres mismos hayan heredado una predisposición a moldear la psicología de sus hijos para que se parezcan más a ellos, para reforzar en sus hijos los rasgos de los padres. Tal predisposición amplificaría las predisposiciones psicológicas heredadas de los hijos con respecto a aquellos rasgos que los padres comparten con ellos, y así, esta predisposición a reforzar y a modelar podría haber sido también positivamente seleccionada, al menos cuando actuó en combinación con otros rasgos psicológicos heredables que sirvieron a la adaptación inclusiva. (Si existiera una predisposición heredable a modelar la psicología de los hijos para que cuadrara con la propia, ¿esperaríamos entonces que su base genética estuviera estrechamente vinculada cromosómicamente a las bases de algunas importantes predisposiciones psicológicas?) No resultaría sorprendente, pues, que las influencias naturaleza-educación llegaran a estar tan inextricablemente unidas que fueran muy difíciles de desenmarañar si se diera el caso de que la naturaleza llevara a los padres a educar para amplificar aquellos rasgos psicológicos que sus hijos comparten con ellos, y a veces, han heredado de ellos.

nes, podemos entonces permitir que los valores heredados y los supuestos de la sociedad resulten determinantes.)

La primera vía parece indebidamente optimista, si no dogmática; la segunda, parece el curso propio de la racionalidad. La segunda es la que me gustaría defender. Sin embargo, considerar todas y cada una de las opciones es ineficiente. Si el sesgo en la presentación social de las razones representara una ponderación de las razones que coincidiera con la que éstas deberían tener, con la ponderación a la que llegaría una persona tras una minuciosa y pausada consideración, entonces lo eficiente sería seguir la orientación de la sociedad. ¿Se pueden comparar los sesgos sociales con una asignación bayesiana de probabilidades previas?

Distintos tipos de adaptación casan con distintos ritmos de cambio. Es ineficiente para cada persona tener que aprenderlo todo desde el principio y tener que construir cada bit de conocimiento por sí misma. Nuestra herencia genética, evolucionariamente inculcada, fue moldeada durante larguísimos períodos de tiempo para cuadrar con o responder a constantes mantenidas a lo largo de esos mismos períodos de tiempo. Que existen objetos duraderos de dimensiones medias que se mueven continuamente, que la gravedad ejerce una fuerte atracción continua procedente del centro de la tierra —esas constantes se han mantenido a lo largo de grandes trechos de nuestro pasado evolucionario—. (Una de las constantes de largo plazo es que otras cosas cambian con frecuencia, de manera que estamos moldeados también para tener, como parte de nuestra dotación genética permanente, mecanismos para responder a esos cambios.) Aquello a lo que responden nuestros genes no cambia —hasta ahora— sino a lo largo de eones.

Lo que se adapta a factores que cambian más rápidamente, en unas cuantas generaciones, no en unos cuantos eones, es, según algunos, las tradiciones sociales, las instituciones y las reglas de conducta.[37] Otros factores cambian en una vida —por ejemplo, la pericia laboral que hay que tener en las sociedades modernas (no tradicionales)—. (Este ritmo puede incrementarse a tal punto que las necesidades laborales cambian significativamente en el curso de la vida de una persona, contrariamente a lo esperado.) Hay otras adaptaciones a factores que cambian lentamente en el curso de una vida, por ejemplo, los hábitos de comportamiento y los vínculos personales duraderos. Luego hay factores que cambian frecuentemente, cada día, o de un momento a otro; nuestro aparato perceptivo está presto a atender a cambios que ocurren con respecto a nosotros, y los mecanismos de información social también están prontos a traer-

nos noticias de cambios más distantes. Hay distintos «conjuntos de adaptaciones armonizadas con cambios medioambientales de distintas duraciones».[38]

A veces, lo racional será aceptar algo porque otros en nuestra sociedad lo aceptan. Consideren el mecanismo de creencia que les lleva a ustedes a aceptar lo que (según ven) la mayoría de los demás acepta. Todos somos falibles, de manera que el consenso de muchas personas falibles probablemente sea más atinado que mi propio punto de vista cuando lo que anda en juego es algún asunto al que todos tenemos el mismo acceso. Para una amplia gama de situaciones, la media de una muestra más amplia de observaciones es probablemente más exacta que una observación individual aleatoriamente seleccionada. (Supongamos que las observaciones se distribuyen normalmente alrededor del valor verdadero, o están determinadas por el valor verdadero junto con factores de error aleatorio.) En lo que hace a esos asuntos, pues, ustedes deberían corregirse a sí mismos para acercarse al punto de vista consensuado, a menos que tengan una razón especial para pensar que los demás andan errados y ustedes no —por ejemplo, ellos han sido inducidos al error por una vía a la que ustedes no están sujetos—.* Si los puntos de vista de la mayoría están formados mitológica y supersticiosamente, mientras que los míos se basan en la literatura científica (o en informes sobre ella), cuyos puntos de vista son construidos de un modo más escrupuloso y fiable, entonces puedo tener razones para pensar que el punto de vista minoritario es correcto. La tradición marxista sostuvo que los puntos de vista de los demás estaban moldeados por mecanismos ideológicos de obnubilación y, por lo mismo, presos de la «falsa consciencia»; quienes supieran más sobre esos mecanismos, o hubieran accedido a las teorías que los desenmascaraban, podrían por ende sentirse justificados en el rechazo de un consenso poco fiable.

Nuestra cuestión ahora es si los mismos sesgos pueden tener una función engendrada por algún mecanismo homeostático social. Eso dependerá de la naturaleza de los procesos sociales. En el caso de los supuestos evolucionariamente inculcados —caso de que los haya—, éstos casan lo bastante con los hechos como para dar una ventana a quienes acceden a ellos fácilmente. ¿Acaso los sesgos sociales en la ponderación de la información y de las razones son el resultado de algún mecanismo selectivo que sintoniza con hechos

* ¿Deberíamos, pues, reinterpretar el famoso experimento de Solomon Asch sobre el conformismo social?

importantes o sirve a empeños muy difundidos en la sociedad? Si es así, esto vendría en apoyo del punto de vista conservador, según el cual hay presunción favorable a las instituciones, las tradiciones y los sesgos existentes. Por mi parte, me declaro escéptico. Sólo si este proceso selectivo fuera severo y fuera deseable el criterio de selección que lo orienta, podríamos garantizar su legítimo peso a cualesquiera sesgos que se produjeran.[39]

Una presunción conservadora en favor de las instituciones, las tradiciones y los sesgos que han existido por mucho tiempo es extremadamente implausible en su tenor literal, a menos que fuera severamente restringida. La esclavitud existió por mucho tiempo, y sigue existiendo la subordinación de las mujeres en la sociedad, la intolerancia racial, el abuso de los niños, el incesto, la guerra y la mafia siciliana. Cuánto peso haya que conceder al hecho de que algo haya durado mucho tiempo depende de por qué ha continuado existiendo, de la naturaleza del test selectivo que ha pasado y del criterio incorporado por este test. (Y la naturaleza del medio en el que la selección ha tenido lugar puede haber cambiado relevantemente, de manera que lo que antaño fue adaptativo ahora ya no lo es.) Es posible sobreestimar la severidad de un test, pero también es posible subestimarla. Los marxistas pensaron que el único test que habían pasado la sociedad capitalista y sus instituciones en los últimos tiempos —desde que completó la tarea de liquidar el feudalismo— fue el de servir a los intereses de la clase dominante, de modo que se propusieron acabar con lo que no consiguieron entender. (Viendo el resurgimiento del marxismo en los ambientes académicos humanísticos y culturales, podría decirse que el marxismo repite ahora como farsa lo que antes fue una tragedia.)

Incluso cuando hay una razón conocida para la supervivencia de algo, e incluso cuando algunas funciones cumplidas por ese algo son valiosas, la cuestión de si su mantenimiento es deseable es harina de otro costal. Eso dependerá de si podemos divisar e instituir una alternativa racionalmente creíble que sea mejor y para la que valga la pena correr los riesgos por ella entrañados. Pero una sociedad sólo elegirá la vía del cambio institucional si se ve forzada a hacer uno u otro cambio. De manera que la toma de decisiones de una sociedad *tenderá* a ser conservadora, mas no como resultado de estimar que lo existente es el producto de algún proceso selectivo imbuido de un criterio laudable.

El capitalismo surgió del feudalismo a través de una serie de transformaciones graduales, cada una de las cuales produjo visibles beneficios positivos por el camino (incrementos de productividad,

en el tamaño de la población que podía ser sostenida, etc.), suministrando así razones e incentivos para proseguir. El capitalismo surgió por un método de «escalar cumbres». (Un mecanismo complejo e interconectado como el ojo puede evolucionar si cada componente produce alguna mejora beneficiosa, si cada incremento de sensibilidad respecto de la luz produce algunos beneficios, aunque sean beneficios menores que los plenos beneficios que se derivan del ojo completo.)[40] Por esa senda sólo está garantizada la producción de un óptimo local, pero aun si el capitalismo no fuera más que eso, sería estable (frente a un movimiento que propugnara lo que creyera un óptimo global) si ninguna secuencia de pequeños pasos, siempre progresivos, llevara desde él hasta un (supuesto) óptimo global diferente. Si tal fuera el caso, la sociedad se enfrentaría a un problema en punto a alcanzar el óptimo global.[41] Evidentemente, somos inteligentes y racionales; podemos mirar hacia adelante y decidir cruzar el valle. Cuanto más numerosos y grandes, empero, sean los cambios necesarios para llegar a la estación de destino, tanto más distante estará ésta. ¿Cuán robustas son las razones para pensar que las cosas irán mejor allí? Cuando las cosas andan razonablemente bien en la posición local, este viaje a lo desconocido parecerá demasiado preñado de graves asechanzas, independientemente de lo bien sostenido que esté por la argumentación teórica. (Pues ¿hasta qué punto son buenas nuestras teorías, hasta qué punto es buena nuestra inteligencia de la sociedad?)

De aquí que no resulte sorprendente que los «experimentos» sociales audaces y globales (que entrañan y requieren cambios simultáneos de muchos factores) resulten más probables cuando la situación general es desesperada y donde hay pocos centros independientes capaces de resistirse al cambio. (La revolución marxista tuvo lugar en Rusia, y no, como esperaban los marxistas, en los países capitalistas avanzados, y fue seguida como modelo en partes del mundo económicamente subdesarrollado.)

Se pueden dar dos tipos de razón para confiar en un viaje social. Primero, puede emprenderse pasito a pasito, y cada pasito puede estimarse beneficioso y generar confianza en el siguiente. Pero esto es tanto como decir que aún no hemos llegado a un óptimo local, de manera que este tipo de razón no sirve para el problema de llegar a un óptimo global distante. (La descripción teórica de la sociedad capitalista en la *Riqueza de las naciones* de Adam Smith llegó sólo después de haberse emprendido el viaje y de haberse cosechado algunos resultados positivos. No sólo los pasos pasados estimularon a dar pasos ulteriores; también lo hizo esta teoría. La teoría,

además, se invistió de autoridad convincente porque explicaba los resultados exitosos de los pasos pasados —¿era «sólo teoría»?—.)

Un segundo tipo de razón para emprender un viaje social es que los experimentos a pequeña escala orientados en este sentido han resultado ya exitosos (en la sociedad existente). Sin embargo, si el óptimo global pudiera funcionar bien sólo una vez instituido para la entera sociedad —¿o internacionalmente?—, pero no incrustado en una sociedad de diferente carácter, entonces no sería posible hallar esos éxitos estimulantes. Aun si los experimentos a pequeña escala funcionaran, subsistiría la cuestión de si sus resultados son extrapolables. ¿Funcionarán a gran escala? ¿Funcionarán con todos y cada uno, no sólo con una población especialmente seleccionada de participantes o monitores?

Una sociedad que vaya razonablemente bien a la luz de sus propios criterios no alcanzará, por sí misma, un óptimo global no conectado con ella mediante mejoras locales; no se ensayarán aquí por primera vez experimentos globales. Pero pueden ensayarse en otras partes, y demostrar su viabilidad y su eficacia. Así también, dados los vínculos y la competición económica a escala internacional, un ejemplo externo puede llevar a la gente a dar su apoyo a una gran modificación en su propia sociedad.[42]

Además de los mecanismos biológicos y sociales que moldean nuestro actuar por razones, está también la automodelación personal, los modos en que modificamos y dirigimos nuestra concepción y nuestra utilización de razones para propósitos que *nosotros* elegimos —usando, es verdad, capacidades biológica y socialmente moldeadas (y también los resultados de la modelación personal previa) en este proceso—. Cualesquiera que fueren las funciones originarias de las razones, podemos servirnos de nuestra capacidad de emplear razones para formular nuevas propiedades de las razones y para moldear nuestra utilización de razones con objeto de manifestar esas propiedades. Podemos, esto es, modificar y alterar las funciones de las razones y, por lo tanto, de la racionalidad.

Capítulo 5

LA RACIONALIDAD INSTRUMENTAL Y SUS LÍMITES

¿BASTA LA RACIONALIDAD INSTRUMENTAL?

¿Basta para qué? replicará el instrumentalista, complacido de que, con la pregunta misma, parezcamos dar por sentado su punto de vista. Pero la cuestión es si la racionalidad instrumental es *toda* la racionalidad.

La noción instrumental de racionalidad puede formularse en términos de teoría de la decisión en el marco de la teoría *causal* de la decisión, cuya noción de conexión causal (probabilista) capta la idea central de racionalidad instrumental: la conexión medios-fines. Debemos (también la teoría causal de la decisión) construir esa conexión laxamente para poder incluir la relación entre una acción y el conjunto más amplio de acciones, del que ella es un caso, un *modo* más —como volar a Chicago es un modo de viajar a Chicago—. Entre los objetivos que hay que conseguir puede contarse el de la realización de una acción, y una manera de conseguirlo es realizar esa acción de un cierto modo; la racionalidad instrumental, que consiste en la consecución efectiva y eficiente de objetivos, tiene que incluir eso. De manera que la racionalidad instrumental no siempre tiene que ver con la consecución de alguna *otra* cosa, de alguna cosa completamente distinta. Y la racionalidad instrumental puede reconocer que el realizar una acción (o el haberla realizado) podría por sí mismo tener algún valor, un valor que (acaso con cierto esfuerzo) podría llegar a incluirse en lo que la acción misma produce. (En lo que sigue, cuando hablo de «causal» y de «instrumental» me refiero al aspecto ejemplificador de, y a la utilidad que reporta, la realización misma de la acción.)

La noción de racionalidad instrumental es una noción poderosa y natural. Aunque se han ofrecido descripciones más amplias de la racionalidad, cualquier descripción que se pretenda completa incluirá la racionalidad instrumental. La racionalidad instrumental se halla en la intersección de todas las teorías de la racionalidad (acaso sea lo único que se halla en esa intersección). En este sentido, la racionalidad instrumental es la teoría del defecto, la teoría que todos

quienes discuten sobre racionalidad pueden dar por sentada, sea lo que sea lo que piensen de ella. Pero creo que hay algo más. La teoría instrumental de la racionalidad no parece necesitar justificación; cualquier otra teoría, sí. Cualquier otra teoría produce razones para poder concluir que aquello demarcado por ella forma realmente parte de la racionalidad. La racionalidad instrumental es la base de partida. La cuestión es si es la racionalidad *toda.*

Algunos objetarán que cualquier extensión de la racionalidad instrumental carece de justificación y que justificar esa extensión sirviéndose de procedimientos que no sean puramente instrumentales no es sino una petición de principios. (Si los procedimientos no instrumentales se justifican con procedimientos instrumentales, ¿quedarán empañados para siempre por la instrumentalidad de su origen?) De modo que podemos preguntarnos por qué deberíamos ser *instrumentalmente* racionales. ¿Por qué habría de perseguir nadie sus deseos y objetivos del modo más efectivo y eficiente? Porque es la manera más probable de satisfacer sus deseos o de lograr sus objetivos al menor coste (consiguiendo y satisfaciendo así la mayor cantidad posible de objetivos y deseos). Mas ¿por qué habría de conseguir sus objetivos y satisfacer sus deseos? Porque esto es lo que quiere. Mas ¿por qué habría alguien de satisfacer *este* deseo? ¿Hay alguna respuesta que no sea circular, alguna respuesta que no haga petición de principios a la hora de justificar la racionalidad instrumental?

Si otros modos de racionalidad no pueden justificarse a sí mismos sin circularidad, tampoco la racionalidad instrumental puede hacerlo.* De manera que, por sí misma, ésta no es una crítica concluyente de los otros modos. En cualquier caso, no está clara la base del rechazo de la circularidad por parte de la racionalidad instrumental. ¿Hay evidencia empírica de que las justificaciones circulares que damos, en situaciones que realmente nos fuerzan a darlas, nos lleven a un grado de satisfacción de los deseos (o de consecución de la verdad) menor que el grado al que nos llevaría algún otro procedimiento (particular)?

No he planteado la cuestión de por qué deberían satisfacerse los deseos con el propósito de sostener que la racionalidad instrumen-

* Los argumentos a favor de la racionalidad instrumental aceptarán presumiblemente criterios que resulten efectivos en punto a conseguir objetivos cognitivos. Sin embargo, el instrumentalista consecuente no puede pensar en exigir ningún requisito racional para la aceptación o la búsqueda de esos objetivos; de manera que, si él está en lo cierto, sus argumentos convencerán (racionalmente) *a lo sumo* sólo a quienes ya compartan esos objetivos.

tal no puede (en principio) constituir el único contenido de la racionalidad. El instrumentalista puede considerar que el objetivo de satisfacer los deseos es un objetivo que nos viene dado, no un objetivo racional, y el crítico que exija que se demuestre la racionalidad de la satisfacción misma de deseos estará importando un criterio que el instrumentalista no tiene por qué aceptar. Tampoco he planteado la cuestión de la justificación de la racionalidad instrumental con el propósito de cuestionar su legitimidad. (No albergo dudas al respecto, al menos en lo que hace a la racionalidad dirigida a objetivos racionales.) Pero sí sostendré que hay también otros modos legítimos de racionalidad, y por lo tanto, que el concepto de racionalidad no se agota en la instrumental. Planteo la cuestión de la justificación de la racionalidad instrumental para prevenir de entrada la objeción instrumentalista de que no pueden justificarse sin circularidad otros modos de racionalidad. Todos están en el mismo barco.

En un enfoque causal-instrumental de la racionalidad, nuestros criterios de racionalidad tienen que depender de nuestra concepción de la naturaleza de este mundo y de nuestra concepción de lo que somos, de nuestras capacidades, facultades, impericias y debilidades. (En un enfoque más amplio, no meramente instrumental, se dará también, aunque sólo parcialmente, esa dependencia.) Lo cierto es que si la racionalidad viene fijada por los efectos, lo que realmente (o generalmente) haya de tener ciertos efectos es una cuestión empírica. Algunos criterios resultarán causalmente más eficaces en punto a conseguir un objetivo en un tipo de mundo que entrañe un tipo de persona, pero no en otro, en el que resultarán más eficaces otros criterios. Es claro que hay una interacción. Usamos los criterios en un momento dado para descubrir el carácter del mundo y nuestro propio carácter, y fundándonos en la inteligencia a la que llegamos modificamos o alteramos nuestros criterios para hacerlos (lo más probablemente posible) más eficaces en este tipo de mundo (según inteligimos ahora el nuestro). El proceso continúa, pues esos criterios nuevos llevan a ulteriores modificaciones en nuestra concepción del mundo y de nosotros mismos, y por lo tanto, a nuevos criterios, y así sucesivamente. Nuestra concepción del mundo y de nosotros mismos, y nuestra noción de lo que haya que considerar racional, andan en continua interacción.

Podríamos pensar en la teoría *pura* de la racionalidad como en la teoría de los criterios que ha de aceptar cualquier tipo de ser en cualquier tipo de mundo, es decir, que han de aceptar todos los seres en todos los mundos. Pero parece improbable que, de *existir* tales criterios, vayan a tener mucho contenido o a llevarnos muy lejos.

No obstante, situados nosotros en un mundo, podrían proporcionarnos un punto de partida para el proyecto de usar esos criterios puros para construir una concepción de ese mundo y de nosotros mismos, modificando entonces correspondientemente los criterios, construyendo una nueva concepción fundada en esos criterios modificados, y así sucesivamente. Pero no hay ninguna razón para creer que esos criterios puros anden por detrás de la historia real de nuestros propios criterios, más plenos de contenido. Nuestros criterios nacieron con impurezas, y el intento filosófico de redimirnos de ese pecado original no ha tenido éxito hasta ahora.

Todavía hay que mencionar una interacción adicional. Nuestros principios de decisión y nuestros principios de razonamiento están interimbricados. Razonamos acerca de los principios de decisión que hay que seguir —el segundo capítulo de este libro constituye un ejemplo de ello—. Y también podemos decidir qué principios de decisión seguir. La política de seguir un particular grupo de principios de razonamiento es un curso de acción. Dos de esos cursos de acción que tengan que ver con distintos grupos de principios de razonamiento podrán entonces evaluarse mediante un principio de decisión que determine qué curso de acción es el mejor. Qué principios de razonamiento acaben considerándose los mejores dependerá de qué principio de decisión se emplee.

Esto es sin disputa relevante para quienes pretendan justificar principios de razonamiento por su fiabilidad en punto a obtener verdades, considerando así que el modo más fiable de razonamiento es también el modo más justificado. También querríamos estimar cómo nos las arreglamos cuando las cosas van *mal* con el principio de razonamiento, cuando éste no lleva a la verdad. Si un principio es más fiable en general, pero desastroso cuando se equivoca, mientras que otro es menos fiable, pero también menos desastroso cuando se equivoca, bien podríamos favorecer el último, renunciando así a cierta fiabilidad a cambio de otros beneficios. Y algunos tipos de verdades pueden resultar más valiosos para nosotros que para otros por razones intelectuales o personales, de manera que podemos vernos inclinados a favorecer un método que logre *este* tipo de verdades, aun siendo la fiabilidad general de este método inferior a la de otro método. Así, pues, un método de razonamiento puede estar él mismo sujeto, en términos de teoría de la decisión, a consideraciones de utilidad esperada, no solamente de probabilidades. También podrían entrar en juego otras consideraciones que llevaran a pensar que la simple maximización de la utilidad esperada resulta inadecuada para la elección de principios de razonamiento, sugiriendo así el uso de un principio de decisión diferente.

Un principio de decisión se establece (tentativamente) mediante un método de razonamiento; un método de razonamiento, mediante un principio de decisión. Un proyecto para investigar todos los posibles pares de principios de decisión y principios de razonamiento que tratara de averiguar cuáles de esos pares, y en qué condiciones, están compuestos por elementos que se dan apoyo mutuo sería un proyecto desmesuradamente ambicioso. Pero deberíamos esperar que en nuestra presente situación se dé un apoyo mutuo significativo entre el conjunto de principios de decisión que consideramos mejor, cualquiera que sea, y el conjunto de principios de razonamiento que hallamos más convincente, cualquiera que sea. Una discrepancia en este punto no sería sino una invitación al cambio. Lo que queremos es que, con el tiempo, se acabe dando una convergencia que podemos llamar racionalmente autocorrección.

Aun si la racionalidad fuera entendida y explicada sólo como racionalidad instrumental, esa racionalidad puede llegar a ser estimada, parcialmente, por sí misma —recuérdense los medios que acaban convirtiéndose en fines de los que nos habló John Dewey—, y así, llegar a adquirir valor *intrínseco*. La *naturaleza* de esa racionalidad sería, vamos a suponer, enteramente instrumental, pero no su *valor*. Valoramos a una persona que crea y decida racionalmente de un modo sensible al balance neto de razones, y pensamos que eso es bueno y admirable por sí mismo, acaso porque decidir y creer así conlleva el uso y la expresión de nuestras más refinadas capacidades, o acaso porque incorpora una integridad admirable y de principio en la guía de nuestras creencias y nuestras acciones mediante razones, no mediante los caprichos y los deseos del momento. Tenemos además el tema tan resaltado por Heidegger: las herramientas instrumentales, usadas con la frecuencia suficiente, pueden llegar a ser extensiones de nosotros mismos; nuestras fronteras pueden extenderse, a su través, hasta los *fines* de esas herramientas a medida que interactuamos con el mundo. Así, la racionalidad —también los principios, el tópico de nuestro primer capítulo—, al comienzo completamente instrumental, puede, si se utiliza suficientemente, llegar a ser una extensión de nosotros mismos y asimilarse a nosotros como una parte importante de nuestra identidad y de nuestro ser.

También podríamos concebir la racionalidad como nuestra particular ruta hacia la *comprensión*, no simplemente como un medio para llegar a ella, sino como un constituyente o componente de la comprensión. Algunos sistemas complejos pueden entenderse —por nuestra parte, al menos— sólo por medio de una teoría articulada, una teoría de la que conocemos sus interconexiones con otras teo-

rías y las razones que vienen en su apoyo, una teoría cuya capacidad para resistir a las objeciones (y cuya disposición a resolverlas) nos resulta familiar. Obsérvese que la racionalidad no es simplemente instrumental en punto a descubrir una teoría así, sino que es también un componente definitorio de lo que haya de significar el comprender esta teoría y los fenómenos que describe. Si la comprensión es algo que ahora valoramos parcialmente por sí misma —y si la racionalidad entra como ingrediente en la naturaleza de tal comprensión—, entonces también la racionalidad ha de valorarse, en parte, por sí misma, intrínsecamente.

Al formular una teoría de la decisión que incorpora, además de la utilidad causalmente esperada, la utilidad simbólica y evidencialmente esperada, empero, hemos ampliado la noción de racionalidad más allá de la simple instrumentalidad. La racionalidad no tiene que ver sólo con lo que (probablemente) se generará o se producirá. En el capítulo 2 llegamos a la conclusión de que la teoría causal de la decisión no es, por sí misma, una teoría plenamente adecuada de la decisión racional. Una decisión racional, dijimos, maximizará el valor decisional, que no es sino la suma ponderada de su utilidad causal, evidencial y simbólica. Pero la racionalidad instrumental queda completamente captada y agotada por la noción de utilidad causal esperada. Puesto que la utilidad causal esperada es sólo *uno* de los aspectos de la racionalidad, tiene que haber más en la racionalidad que la mera instrumentalidad. La gente siempre ha pensado así, obvio es decirlo. Los factores evidenciales y simbólicos siempre han funcionado con consecuencias sociales *muy* significativas en la historia humana (recuérdese, una vez más, la literatura sobre el papel desempeñado en el desarrollo del capitalismo por la idea calvinista de los *signos* de ser elegido).

Ya hemos incluido estos factores no instrumentales en nuestra teoría de la creencia racional en dos etapas. La primera etapa elimina como candidatos para la creencia aquellos enunciados cuyo valor de credibilidad es menor que la de otros enunciados competitivos. Esos valores de credibilidad serán determinados por vínculos reticulares acordes con conexiones (probabilísticas) fácticas, de manera que los pesos transferidos por esos valores serán instrumentales respecto de la consecución de creencias verdaderas o de otros objetivos cognitivos. Pero la segunda etapa, en la que se decide si hay que creer un enunciado cuyo valor de credibilidad no se ve superado por el de ningún enunciado competitivo, es explícitamente no instrumental. Lo que ha de juzgarse es el valor decisional de creer el enunciado candidato, es decir, su ponderación de utilidades evi-

denciales, simbólicas y causales. (El instrumentalista usaría la regla 2, interpretada por la teoría causal de la decisión, no la regla 5.)

He partido del supuesto de que cuando está en juego el valor decisional, y no simplemente la utilidad causalmente esperada, entonces el cálculo no es enteramente instrumental. Pues nos preocupamos entonces por los resultados producidos, por sus probabilidades de ser producidos, pero también por aquello a que indiciariamente se apunta y es simbolizado. No obstante, ¿podría el instrumentalista sostener que también esto no es sino cálculo instrumental? Si planteamos que el objetivo es el valor decisional máximo, y si hacemos que la función relevante de utilidad varíe linealmente con ello, ¿no estaría la persona eligiendo la acción causalmente más efectiva para producir un máximo de valor decisional, actuando, así pues, de nuevo instrumentalmente? Parece que aquí no todas las probabilidades relevantes serán causales; la relación de la acción con su utilidad simbólica o con su utilidad evidencial no será causal. Con todo, nuestra descripción ampliada de la racionalidad instrumental incluye ya el modo en que una acción constituye un modo de hacer otra acción. Entonces ¿por qué no incluir en la instrumentalidad esas relaciones en que se halla una acción con su utilidad evidencial y simbólica? Esto convertiría en trivial a la cuestión de la instrumentalidad, pero aún subsistiría un punto interesante, que ahora, empero, habría de redescribirse. El objetivo de algunas acciones sería ahora el máximo valor decisional, que es no sólo un objetivo no instrumental —siempre hemos sabido que los objetivos de las acciones instrumentales pueden ser no instrumentales—, sino también un objetivo que describe y prescribe una acción que se halla en una relación no instrumental con un objetivo. La noción amplia, pues, sólo consigue «salvar» la instrumentalidad entendiendo como objetivo de una acción «instrumental» el que esa acción esté en relaciones no instrumentales con otros objetivos. Y para lo que aquí nos interesa, esto es lo que constituye la *no*-instrumentalidad.

¿Qué ocurre, empero, con el valor de credibilidad de un enunciado, determinado como éste está por una red procesadora de razones a favor y en contra? ¿Es el valor de credibilidad completamente instrumental, en el sentido de que los pesos resultantes en la red estarán únicamente determinados por la efectividad de los mismos en punto a conseguir objetivos cognitivos varios, tales como la creencia verdadera, la potencia explicativa y la simplicidad? Eso dependerá de la naturaleza de la regla de retroalimentación, de la regla de aprendizaje ínsita en la red del sistema de procesamiento que fija los valores de credibilidad. ¿Qué es lo que retroalimenta, y de acuer-

do con qué reglas de revisión lo hace? Habrá margen aquí para factores evidenciales y simbólicos?[1]

Mi argumento de que la racionalidad instrumental no es toda nuestra racionalidad no era desinteresado. Si decimos que los seres humanos son simplemente seres humanos, parece que estamos rebajando nuestra estatura. El hombre es el único animal que no se conforma con ser sólo un animal. (Puesto que mi argumento está motivado, ustedes —y yo también— deberíamos estar alerta para corregir los posibles sesgos del argumento en su tratamiento de las razones.) Es simbólicamente importante para nosotros que no todas nuestras actividades se propongan satisfacer nuestros deseos dados. Los principios, como hemos tenido ocasión de ver, proporcionan uno de los medios para controlar y remodelar nuestros deseos. (Kant les pidió demasiado, empero, al divorciarlos del deseo y esperar de ellos que generaran acciones fundadas únicamente en el respeto del principio mismo.)

Una de las maneras que tenemos de no ser sólo instrumentalmente racionales es preocuparnos por los significados simbólicos, independientemente de lo que causan o producen. El propugnador de la racionalidad instrumental no puede sostener fácilmente que esta preocupación es irracional, pues carece de criterio de racionalidad relevante para ello —¿por qué habría de ser esta preocupación más irracional que cualquier otra?—. Los significados simbólicos son un modo de colocarnos a nosotros mismos por encima del nexo causal de los deseos, y es simbólicamente importante para nosotros el hacerlo. ¿Significa esto que actuar a partir de significados simbólicos, tomar en cuenta las utilidades simbólicas, es un *medio* para colocarnos por encima de este nexo humano? —preguntará con una sonrisa el instrumentalista—. Quizá, mas aun si no consiguiera eso, podría *simbolizarlo.** Incluso respecto de los procesos de formación y mantenimiento de nuestras creencias, pues, podemos preocuparnos no simplemente de lo que produzcan causalmente esos procesos, sino también de lo que simbolizan. Nuestra discusión de los principios en el primer capítulo era, en parte, instrumental; consideramos

* Puesto que orientarse hacia el significado simbólico tiene un significado simbólico para nosotros, aparte de sus consecuencias reales, la utilidad simbólica puede venir en apoyo de sí misma. La acción misma de secundar el principio de tomar en cuenta la utilidad simbólica tiene utilidad simbólica, convirtiéndose así en uno de los casos cubiertos por el principio. También la racionalidad instrumental puede venir en apoyo de sí misma y subsumirse en sí misma. Sobre la autosubsumción, véase mi *Philosophical Explanations* (Cambridge, Mass.: Harvard University Press, 1981), págs. 119-121, 131-140.

las funciones a las que podían servir los principios. Ahora vemos, en cambio, una posible metafunción —colocarnos por encima del servicio a otras funciones—, de modo que secundar principios puede tener también una utilidad simbólica.

He sostenido que hay modos legítimos de racionalidad además del instrumental —los evidenciales y los simbólicos—, pero subsiste la cuestión de la primacía que puedan disputarse los distintos modos legítimos. Y la de qué orden de prioridades decidiremos establecer.

LAS PREFERENCIAS RACIONALES

Hay, obvio es decirlo, una crítica corriente a la noción de racionalidad instrumental, y es que ésta trata de agotar el entero dominio de la racionalidad. Algo es instrumentalmente racional con respecto a objetivos, fines, deseos y utilidades dados cuando es causalmente efectivo en la realización de los mismos. Pero la noción de racionalidad instrumental no nos proporciona ninguna manera de evaluar la racionalidad de esos objetivos, fines y deseos, salvo en el caso de que sean instrumentalmente efectivos en punto a realizar ulteriores objetivos que se toman como dados. Ni siquiera para objetivos cognitivos que se toman como dados. Ni siquiera para objetivos cognitivos tales como creer la verdad parecemos tener otra justificación que la instrumental. Hasta el presente no tenemos ninguna teoría adecuada de la racionalidad substantiva de objetivos y deseos que pueda replicar concluyentemente a la aseveración de Hume: «No es contrario a la razón preferir la destrucción del mundo entero a un rasguño en mi dedo».[2]

Me propongo dar algunos pasos tentativos en la dirección de una teoría de la racionalidad substantiva de deseos y objetivos. Permítaseme avanzar que no trato de que resulten aceptables las *particulares* condiciones que propondré, ni menos de defender sus particulares detalles. Espero mostrar más bien cuán prometedor es el espacio abierto a condiciones como las que discutiré, y qué direcciones pueden seguirse para ir más allá de Hume.

Como filósofo, fiel a la forma, empezaré una discusión de contenido atendiendo a la forma. Un pasito más allá de Hume, pero nada que él se viera en la necesidad de detener, según pienso, están las restricciones al modo en que las preferencias dependen unas de otras, restricciones formuladas en las condiciones estándar de Von Neumann-Morgenstern, o en las variantes de ellas que presenta la

teoría de la decisión —por ejemplo, que las preferencias sean transitivas, que cuando dos opciones traigan consigo dos posibles consecuencias idénticas, pero difieran en las probabilidades que les asignan, resultará preferida la opción que confiere la mayor probabilidad a la consecuencia preferida.[3] Algunas de estas condiciones se justifican por consideraciones instrumentales tales como el argumento del «bombeo de dinero», que concluye que las preferencias son transitivas,* mientras que otros se presentan como normativamente atractivos en su mismo tenor literal. (A menos que pueda darse también a estos últimos una justificación instrumental, ¿no constituye esto ya un paso más allá de la racionalidad instrumental?) La teoría contemporánea de la decisión da este paso más allá de Hume: aunque no dice que alguna preferencia, individualmente considerada, sea irracional, sí dice que un grupo de ellas puede serlo tomado en conjunto. Supongamos que hay principios normativos que definen la estructura de varias preferencias tomadas en conjunto, y que esos principios son condiciones de racionalidad. (La literatura pertinente está llena de contraejemplos y objeciones putativos a algunas de las condiciones de Von Neumann-Morgenstern; se trata aquí de usar no estas particulares condiciones, sino algún conjunto apropiado de condiciones.)[4]

> **I.** La persona satisface las condiciones de Von Neumann-Morgenstern, o algún otro conjunto apropiado de condiciones, respecto de las preferencias y sus relaciones con las probabilidades.

* La idea es que con preferencias no transitivas, por ejemplo, preferir *x* a *y*, *y* a *z*, y *z* a *x*, una persona que empiece con *z* puede verse conducida a pagar una pequeña cantidad para mejorar su situación y conseguir la *y* que prefiere, otra pequeña cantidad para conseguir la *x* que prefiere a *y*, y aun otra pequeña cantidad para conseguir la *z* —la *z* con la que empezó—, que prefiere a *x*, acabando el ciclo como un perdedor neto. Véase Donald Davidson, J. McKinsey y Patrick Suppes, «Outlines of a Formal Theory of Value», *Philosophy of Science* 22 (1955): 140-160, en donde se atribuye el argumento a Norman Dalkey. El argumento parte del supuesto de que una persona quiere actuar siempre fundándose en cada preferencia individual, aisladamente considerada, y quiere actuar fundándose *repetidamente* en cada una de sus preferencias, independientemente de lo que sepa acerca del modo en que casan unas con otras conjuntamente consideradas, independientemente de si predice o no que la secuencia de sus acciones fundadas en preferencias individuales aisladas le habría de llevar a este tipo de engorros. Se trata, obviamente, de un supuesto implausible. Puesto que el argumento del bombeo de dinero está concebido para justificar la condición normativa de que las preferencias deben ser transitivas, sería interesante hallar una formulación exacta de la condición normativa de la que ese argumento depende, y a la que presupone.

Esto sugiere al menos una condición adicional que deben satisfacer las preferencias de una persona para que ésta sea racional, a saber, que debe preferir satisfacer las condiciones normativas a no satisfacerlas. En realidad, para cada condición estructural válida *C* de racionalidad, ya se trate de racionalidad de las preferencias, de las acciones o de las creencias:

> **II.** La persona prefiere satisfacer la condición de racionalidad *C* a no satisfacer *C*.[5]

(Esta condición debería formularse como una condición *prima facie* o con una cláusula *ceteris paribus*, lo mismo que varias otras que se declaran más abajo. La persona que sabe que será asesinada si satisface siempre la condición de que la indiferencia sea transitiva, o la condición de no creer ningún enunciado cuyo valor de credibilidad sea menor que el de un enunciado incompatible, puede preferir no satisfacerlas.) Puesto que hay que suponer que la persona será instrumentalmente racional,

> **III.** La persona deseará, permaneciendo iguales las demás cosas, los medios y las precondiciones para satisfacer las condiciones *C* de racionalidad.

Esas condiciones *C* de racionalidad no sólo tienen que ver con la estructura de las preferencias, sino que incluyen también las condiciones estructurales apropiadas de racionalidad, cualesquiera que sean. De aquí que la persona desee los medios y las precondiciones de la creencia racional, que desee los medios y las precondiciones necesarias para la efectiva asignación de valores de credibilidad (y para decidir sobre la utilidad de albergar una creencia particular).

Una persona carece de integración racional cuando prefiere alguna alternativa *x* a otra alternativa *y*, y sin embargo prefiere no tener esta preferencia, esto es, cuando también prefiere no preferir *x* a *y* a preferir *x* a *y*. Cuando una preferencia de segundo orden de este tipo entra en conflicto con una preferencia de primer orden, está abierta la cuestión de cuál de esas preferencias debería ser cambiada. Lo que *está* claro es que no casan bien entre sí, y una persona racional *preferiría* que esta situación no se diera (o no perdurara).[6] Así, pues, nos deslizamos hacia una exigencia de que la persona genere un tercer orden de preferencias, a saber: preferir que no se dé el conflicto de preferencias. Sea *S* la situación conflictiva en cuestión, en la que la persona prefiere *x* a *y*, y sin embargo, prefiere no

tener esa preferencia, es decir, sea *S*: *x*P*y* & [no-(*x*P*y*) P (*x*P*y*)]. Entonces,

> **IV.** Para cada *x* e *y*, la persona prefiere no-*S* a *S*, permaneciendo iguales las demás cosas.

Eso no significa que la persona deba elegir no-*S* antes que *S*, pase lo que pase. Un adicto que desee no desear heroína puede saber que no tiene modo factible de anular su deseo de primer orden de heroína, y así, saber que el único modo de resolver el conflicto consiste en eliminar su deseo de segundo orden de no tener el deseo de primer orden. Ello no obstante, puede preferir mantener el conflicto entre deseos porque, con él, la adicción no será tan consumadamente secundada, ni estará su deseo adictivo tan cargado de tacha.[7]

Hume sostiene que todas las preferencias son igualmente racionales. Pero una buena inteligencia de lo que son las preferencias, y de aquello para lo que sirven, podría dejar la vía expedita para imponer condiciones adicionales. En teorías recientes, una preferencia se entiende como una disposición a elegir una cosa antes que otra.[8] La función de las preferencias, la razón de que la evolución nos inculcara la capacidad de tenerlas, es desembocar en la elección preferente. Pero sólo se pueden realizar elecciones preferentes en algunas situaciones; cuando se está vivo, cuando se tiene capacidad para conocer las alternativas, cuando se tiene capacidad para realizar una elección, cuando se puede ejecutar una acción en el sentido de una alternativa elegida, cuando no hay interferencias en esas capacidades que hagan imposible su ejercicio. Son éstas precondiciones (medios) de la elección preferencial. Ello es que no hay por qué preferir que estas condiciones perduren; algunos tienen razones para preferir estar muertos. Pero necesitan una razón creo; la mera preferencia por estar muerto, sin razón alguna, es irracional. Hay una *presunción* de que la persona preferirá que sean satisfechas las condiciones necesarias para la elección preferencial, las condiciones necesarias para que pueda darse alguna elección preferencial; no es necesario que tenga realmente esa preferencia, pero sí necesita una razón para no tenerla.

> **V.** A falta de una razón particular para no preferirlo, la persona prefiere que todas y cada una de las precondiciones (medios) para poder hacer alguna elección preferencial sean satisfechas.

Así, una persona prefiere estar viva y no morir, tener una capacidad para conocer las alternativas y no tener obstruida esa capaci-

dad, tener capacidad para realizar una elección y no ver destruida esa capacidad, y así sucesivamente.[9] Podríamos añadir entonces,

> **VI.** La persona prefiere, manteniéndose todas las demás cosas igual, que las capacidades que constituyen precondiciones para la elección preferencial no se vean interferidas por alguna desgracia (= una alternativa vitanda) que le haga preferir no ejercer nunca esas capacidades en otras situaciones.

Me parece que hay algo más que decir acerca de las razones. (Lo propongo muy tentativamente; se necesita más elaboración para resolver este asunto correctamente.) Supongamos que yo prefiero x a y, sin ninguna razón en absoluto para ello. Entonces, estaré dispuesto, debería estar dispuesto, a invertir mi preferencia si con ello ganara alguna cosa que prefiero tener. Debería estar dispuesto, si en mi poder estuviera, a invertir mi preferencia, a preferir ahora y a x para obtener 25 céntimos más. ¿Y no habría de preferir yo lo segundo a lo primero?* Quizá no, quizá la preferencia de x sobre y es muy *intensa*, sin que haya razón alguna de por medio. Tener una preferencia intensa sin que haya razón alguna de por medio es, según me parece, anómalo. Dado que la tengo, actuaré conforme a ella; pero el maridaje con ella es irracional, es irracional pagar el coste de secundarla o de mantenerla cuando no tengo ninguna razón para ello. O quizá yo prefiero preferir x a y a no tener esta preferencia, y lo prefiero con la intensidad suficiente como para despreciar 25 céntimos. De manera que esta preferencia de segundo orden para preferir x a y podría indisponerme a abandonar esta última preferencia. Mas ¿por qué tengo esta preferencia de segundo orden? Yo diría que, a diferencia de una preferencia de primer orden arbitraria, una preferencia de segundo orden necesita el respaldo de alguna razón. Una preferencia de segundo orden para preferir x a y es irracional a menos que la persona tenga alguna razón para preferir x a y. Es decir, la persona debe tener una razón para preferir tener esta preferencia de primer orden —quizá se la inculcó su madre, o quizá ahora sea esta preferencia parte de su *identidad*, y por lo tanto, algo que no podría desear cambiar—[10] o tener una razón directa para preferir x a y, una razón que tenga que ver con los atributos de x y de y. ¿Pero qué es una razón directa? ¿Debe una razón, en este contexto, ser algo distinto de otra preferencia? Debe al menos ser otra pre-

* Dejemos de lado la consideración de que si alguien me ofrece 25 céntimos, yo podría preferir no ver determinadas mis preferencias por esta fuente externa.

ferencia que funcione como una razón, esto es, una preferencia *general,* aunque rebatible. Suele pensarse que tener una razón para preferir *x* a *y* entraña conocer algún rasgo *F* de *x*, tal que, en general, manteniéndose iguales todas las demás cosas, ustedes preferirían cosas con *F* a cosas sin *F*, entre las cosas del tipo de las *x*.* (Preferir bebidas frías a bebidas calientes no significa preferir habitaciones frías a habitaciones calientes.)

> **VII.** Si la persona prefiere *x* a *y*, o bien: (a) la persona está dispuesta a cambiar y a preferir *y* a *x* a cambio de una pequeña ganancia; *o bien,* (b) la persona tiene alguna razón para preferir *x* a *y*; *o bien* (c) la persona tiene alguna razón para preferir la preferencia de *x* antes que *y* en vez de no preferir esa preferencia.

No digo que *todas* las preferencias de una persona requieran razones —no está claro qué habría que decir sobre las preferencias situadas al más alto nivel; quizás estén ancladas en preferencias situadas más abajo—, pero las preferencias de primer orden andan necesitadas de razones cuando la persona no está dispuesta a cambiarlas. Una vez embarcados en el ámbito de las razones para las preferencias, podemos considerar cómo se relacionan las razones más generales con las razones menos generales, podemos imponer condiciones de consistencia entre las razones, etc. El camino está expedito para exigir ulteriores condiciones normativas a las preferencias, al menos a aquellas preferencias que la persona no está dispuesta a cambiar como se cambia de sombrero. Señaladamente en el caso de las preferencias contrarias a las precondiciones de la elección preferencial antes mencionada, una persona necesitará no sólo razones cualesquiera, sino razones de un cierto peso, lo que significa, por lo menos, que las razones deben estar imbricadas con muchas de las preferencias restantes de la persona, quizás a varios niveles.[11]

Podríamos querer añadir también que los deseos y las preferencias se hallen en equilibrio, en el sentido de que conocer las causas de tenerlos no les llevará a ustedes (a querer) dejar de tenerlos. Los deseos y las preferencias resisten el conocimiento de sus causas.[12]

> **VIII.** Los deseos y las preferencias de la persona están en equilibrio (con las creencias de la persona acerca de sus causas).

* El lector mismo puede derivar lo que sucede cuando, al revés, lo importante es algún rasgo negativo de *y*.

Puesto que las preferencias y los deseos tienen que ser realizados y satisfechos, una persona cuyas preferencias estuvieran estructuradas de tal modo que siempre deseara hallarse en otra situación —prefiriendo *y* a *x* cuando tiene *x* y prefiriendo *x* a *y* cuando tuviera *y*— estaría condenada a la insatisfacción, a más insatisfacción de la que es inherente a la condición humana. No siempre la hierba de otro lugar debería ser más verde. Así que

> **IX.** Para ningún *x* y para ningún *y*, la persona preferirá siempre *x* cuando se dé *y* e *y* cuando se dé *x*. (Esas preferencias condicionales no son preferencias tales que para algunos *x* e *y* la persona prefiere *x* a *y* / dado que se dé *y*, y prefiere *y* a *x* / dado que se dé *x*.)

Los deseos no son meras preferencias. Hay un nivel de filtro o de procesamiento interpuesto en el camino que va de las preferencias a los deseos —lo mismo que (como tendremos ocasión de ver) hay otro en el camino que va de los deseos a los objetivos—. Podríamos decir que los deseos racionales son los deseos de posible cumplimiento, o al menos los que ustedes creen de posible cumplimiento, o al menos los que ustedes no creen de imposible cumplimiento. Seamos lo más cautos posible y digamos que

> **X.** La persona no tiene deseos que sabe de imposible cumplimiento.

Quizá sea correcto preferir volar a pelo, pero no es racional para una persona el *desearlo*. (Podría ser racional, en cambio, desear que fuera posible.) Los deseos, a diferencia de las meras preferencias, han de entrar en algún proceso de decisión. Deben pasar algunos tests de viabilidad, y no simplemente en condiciones de aislamiento: los deseos deben ser coposiblemente satisfacibles, de consuno. Y cuando se descubre que no lo son, deben ser susceptibles de cambio, aunque un deseo alterado o extinguido puede sobrevivir como preferencia.*

Los objetivos, a su vez, son distintos de las preferencias o de los deseos.[13] Tener o aceptar objetivos es usarlos a modo de filtros para dejar fuera de consideración, en las situaciones de elección, aquellas acciones que no sirvan lo bastante bien a esos objetivos, o que

* No me intereso aquí por la palabra *deseo* ni por qué fenómenos cubre realmente esta palabra. Quizá sea el término *objetivos* el que cubra las cosas, la imposibilidad de realización conjunta de las cuales debemos conocer. Lo importante es la distinción conceptual que entraña restricciones crecientes, no el rótulo.

no los sirvan en absoluto. Para seres de limitada capacidad, que no pueden considerar y evaluar en cada momento todas las acciones posibles que están a su disposición —traten de enumerar todas las acciones que ahora mismo están disponibles para ustedes—, un filtro de este tipo es crucial. Por lo demás, podemos usar los objetivos para generar acciones dignas de ser seriamente consideradas, acciones que *sí* sirvan a esos objetivos.[14] Y los objetivos proporcionan dimensiones visibles a los resultados, dimensiones que cobrarán peso en la estimación de la utilidad de esos resultados. Dadas esas múltiples e importantes funciones de los objetivos, podría esperarse que un objetivo importante y estable en el tiempo mereciera que le dedicáramos uno de los pocos canales de alerta de que disponemos con objeto de registrar vías prometedoras para su logro, de controlar el modo en que los estamos haciendo, etc.[15]

¿Cómo surgen nuestros objetivos? ¿Cómo son seleccionados? Resulta plausible pensar que surgen a partir de una matriz de preferencias, deseos y creencias acerca de probabilidades, posibilidades y viabilidades. (Y entonces los objetivos reorganizan nuestros deseos y preferencias, otorgando primacía a algunos e invirtiendo otros, si la inversión de preferencias casa con o promueve el objetivo.)[16] Una posibilidad es que los objetivos surjan por aplicación de la teoría de la utilidad esperada. Para cada objetivo G_i, trata la persecución G_i como una acción con su propia distribución de probabilidades sobre los resultados y computa la utilidad esperada de esa «acción». Adopta el objetivo que traiga consigo la máxima utilidad esperada, y luego úsalo para generar opciones, excluir otras, y así sucesivamente.

Hay una objeción a este modo expedito de hacer casar los objetivos con el marco de la utilidad esperada. El convertir a algo G_i en un objetivo trae consigo dilatados efectos. Ahora, G_i funcionará como un mecanismo exclusionario y tendrá un estatus muy distinto del de otro posible objetivo G_j, que se acercaba mucho a él, pero no conseguía generar una utilidad esperada máxima. Una diferencia marginal marca ahora una gran diferencia.[17] Parecería que grandes diferencias, como las derivadas del hecho de que una cosa cuadre dentro del marco mientras que otras quedan excluidas de él, debería arraigar en diferencias preexistentes significativas.* Consi-

* Al menos cuando no hay necesidad práctica de desviarse de esa norma, como ocurre con la diferencia entre la última persona admitida y la primera persona rechazada en un programa con un número fijo de plazas de admisión, ni necesidad tampoco de incluir muchas entidades marginalmente distintas en un número más pequeño de categorías clasificatorias.

deremos la teoría descriptiva de la decisión propuesta por Henry Montgomery. En ella, un individuo trata de justificar una elección hallando una estructura dominante, y se sirve de mecanismos tales como combinar y alterar atributos y colapsar alternativas en orden a conseguir que una acción domine débilmente a todas las demás en todos los atributos (considerados). De esta guisa consigue evitarse el conflicto, pues una acción resulta claramente la mejor; no hay ninguna razón para hacer otra.[18] ¿Abrirá siempre esa dominación un *hiato* entre las acciones, un hiato lo bastante significativo como para marcar una diferencia cualitativa con efectos de largo alcance, pudiendo así resultar de ayuda en la fijación de los objetivos? Sin embargo, una acción puede dominar débilmente a otra cuando hay seis dimensiones y las dos acciones empatan en cinco de ellas, mientras que en la sexta la primera acción es (sólo) ligeramente mejor. Aun en este marco, parecemos necesitar algo más que la simple dominación; quizá necesitamos que se dé una victoria *fuerte* en una dimensión, o vencer en muchas de ellas.

Volviendo al marco de la utilidad esperada, podríamos decir que hay que elegir el objetivo G_i no simplemente cuando tiene un máximo de utilidad esperada, sino cuando derrota *decisivamente* a los otros candidatos a objetivos. Para cada j, $UE(G_i) - UE(G_j)$ es mayor o igual que alguna cantidad q fija, positiva y definida. (Subsiste, con todo, un problema similar, aunque menor. G_i derrota decisivamente a los otros objetivos, pero no hay diferencias decisivas entre derrotar decisivamente y no hacerlo; la diferencia $UE(G_i) - UE(G_j)$ podría apenas llegar a, o quedarse incluso a un paso de, q.)

Convertir algo en un objetivo consiste, en parte, en adoptar un deseo de encontrar una ruta viable desde donde ustedes se hallan hasta el logro de este objetivo.[19] Por lo tanto,

> **XI.** Una persona no tendrá un objetivo si sabe que no hay ninguna ruta, por larga que sea, que conduzca, desde su presente situación, al logro de este objetivo.

Además, podríamos decir que una persona racional tendrá algunos *objetivos* e investigará las rutas viables que apuntan a ellos, no se limitará a tener preferencias y deseos. Filtrará negativamente acciones que no pueden alcanzar esos objetivos, traerá a consideración acciones que podrían alcanzarlos, etc. Y algunos de esos objetivos tendrán alguna estabilidad, de manera que pueden ser perseguidos a lo largo del tiempo con alguna perspectiva de éxito.

XII. Una persona tendrá algunos objetivos estables.

Una persona racional considerará no sólo resultados particulares (externos), sino también lo que ella misma es, y tendrá algunas preferencias sobre las distintas maneras posibles de ser. Sea *Wp* la manera de ser que una persona cree que le sobrevendrá cuando *p* sea el caso; sea *Wq* la manera de ser que cree que le sobrevendrá cuando *q* sea el caso. (Esto incluye las vías por las que *p* o *q* causarán, o moldearán o impactarán en su manera de ser.) Hay una presunción, que puede ser anulada por razones, de que las preferencias sobre maneras de ser tendrán primacía sobre otras preferencias personales de nivel más bajo. (Las preferencias personales son preferencias derivadas únicamente de la estimación de los beneficios para uno mismo.)

XIII. Si la persona prefiere *Wp* a *Wq*, entonces (manteniéndose iguales todas las demás cosas) no preferirá (personalmente) *q* a *p*.[20]

La condición XIII sostiene que la manera de ser de una persona, el tipo de persona que es, tendrá mayor peso en sus preferencias que (las que de otro modo serían) sus preferencias personales. (¿Está culturalmente determinada esta condición, resulta sólo plausible para personas pertenecientes a ciertos tipos de culturas?)

El argumento del libro holandés, según el cual las creencias sobre probabilidades de alguien deberían satisfacer los axiomas de la teoría de la probabilidad, dice que si no los satisfacen, y si ese alguien está siempre dispuesto a apostar de acuerdo con estas creencias sobre probabilidades, entonces un tercero puede disponer las cosas de tal modo que perderá sin lugar a dudas dinero y acabará, por tanto, en la alternativa menos preferida. Este argumento dice que si sus creencias (probabilísticas) son irracionales, puede estar seguro de que acabará peor en su escala de utilidad. Podríamos buscar un doble de este argumento, imponiendo la condición:

XIV. Los deseos de una persona no están constituidos de tal forma que el actuar a partir de ellos garantice que la persona acabará teniendo creencias o probabilidades irracionales.[21]

Varias cosas podrían acabar siendo barridas por esta condición: desear creer algo sin importar la evidencia; desear pasar el tiempo con un conocido mentiroso sin tomar precauciones; desear colocarse uno mismo —con alcohol, drogas o algo por el estilo— en una

situación que habrá de tener efectos *perdurables* en la racionalidad de las propias creencias. Mas, formulado como está, este requisito es demasiado fuerte; quizás actuar según el deseo le reporte a la persona algo que (legítimamente) valora más que evitar determinadas creencias o probabilidades irracionales.[22] Análogamente, la exigencia del libro holandés resulta demasiado fuerte en su formulación habitual. Pues no es imposible que en el mundo se dé alguna situación en la que tener probabilidades incoherentes traiga consigo un beneficio mayor —alguien podría concederles a ustedes un gran premio por tener probabilidades incoherentes— que la pérdida generada por las apuestas. El argumento del libro holandés dice que la pérdida está garantizada, pero aún podría ser contracompensada; de manera que las creencias o las probabilidades irracionales que ustedes pensaban que violaban la condición XIV podrían ser contracompensadas. Para evitar esa situación es necesario relajar la formulación del argumento del libro holandés, lo mismo que la condición XIV.

Esas catorce condiciones nos sitúan ya, a considerable distancia de Hume, en el camino de las restricciones substantivas a las preferencias y a los deseos. La información empírica acerca de las precondiciones reales de satisfacción de las condiciones de racionalidad y de realización de elecciones preferenciales —exigidas por las condiciones III, V y VI— podría requerir un contenido substantivo bastante concreto a las preferencias y a los deseos, y tanto más concreto cuanto que combinado con las restricciones puestas por las otras condiciones.

¿Se puede seguir por esta vía y proceder a una ulterior definición del contenido? Una ruta prometedora podría tratar de emular en el caso del deseo lo que hacemos con la racionalidad de las creencias. Por ejemplo, se ha dicho que una creencia es racional si está racionalmente formada por un proceso fiable cuyas operaciones dan lugar a un alto porcentaje de creencias verdaderas. Es verdad que los detalles son más complicados, pero podríamos esperar hallar paralelismos también para estas complicaciones. Un deseo racional, pues, sería un deseo formado por un proceso fiable cuyas operaciones dieran lugar a un alto porcentaje de ____ deseos. Mas ¿cómo llenar el espacio en blanco? ¿Qué corresponde, en el caso de los deseos, a la verdad de las creencias? Por el momento, no tengo ningún criterio substantivo que proponer.

Pero podemos usar nuestras condiciones anteriores, y cualesquiera condiciones similares, para definir el objetivo de este proceso: un deseo o preferencia que es racional sólo si está formado por un

proceso que fiablemente da lugar a deseos y preferencias que satisfacen las condiciones anteriores sobre cómo han de estar estructuradas las preferencias, a saber, las condiciones I-XIV. Esto dice más que la mera afirmación de que estas catorce condiciones han de ser satisfechas, pues cualquier proceso (que podamos seguir) que fiablemente consiga la satisfacción de estas condiciones puede restringir ulteriormente los deseos y las preferencias de una persona.

XV. Una preferencia o un deseo particular es racional sólo si hay un proceso *P* para llegar a preferencias y deseos, y
(a) se llegó a esa preferencia o deseo a través de este proceso *P*, y
(b) este proceso *P* da fiablemente lugar a deseos y preferencias que satisfacen las condiciones normativas estructurales I-XIV, y
(c) no hay un proceso más restricto *P'*, tal que se llegara al deseo o a la preferencia a través de *P'*, y *P'* tendiera a producir deseos y preferencias que no satisfacen las condiciones I-XIV.*

Si decimos que las preferencias y los deseos son *racionalmente coherentes* cuando satisfacen las condiciones I-XIV (y condiciones similares), entonces la condición XV dice que una preferencia o un deseo es racional sólo si (es racionalmente coherente y) se accede a él *mediante un proceso* que da lugar a preferencias y deseos racionalmente coherentes.

No sólo puede este proceso *P* dar lugar a preferencias y deseos racionalmente coherentes; también puede *pretender* tales preferencias y deseos, puede modelar y orientar las preferencias y los deseos en el sentido de la coherencia racional. El proceso *P* puede ser un mecanismo homeostático, un mecanismo uno de cuyos estados-objetivo es que las preferencias y los deseos sean racionalmente coherentes. En tal caso, una *función* de las preferencias y de los deseos será ser racionalmente coherentes. (Análogamente, si el mecanismo *C* de formación de creencias pretende creencias que sean aproximadamente verdaderas, entonces una de las funciones de las creencias será ser aproximadamente verdaderas.)

Podríamos por lo tanto añadir nuestra siguiente condición.

* La cláusula c es una condición pensada para lidiar con el problema de la clase de referencia y para excluir deseos formables también a través de un subproceso que fiablemente da lugar a violaciones de las condiciones normativas más arriba expuestas. No debería concederse ningún peso a los detalles particulares de la cláusula c tal como está formulada aquí; se trata simplemente de reconocer el problema y de reservar el lugar para una condición más satisfactoria que pueda lidiar con él.

XVI. El proceso *P* que da lugar a preferencias y deseos pretende que éstos sean racionalmente coherentes; es un mecanismo homeostático, un mecanismo uno de cuyos estados-objetivo es que las preferencias y los deseos sean racionalmente coherentes.

Y, análogamente,

XVII. El mecanismo cognitivo *C* que da lugar a creencias pretende que esas creencias satisfagan objetivos cognitivos determinados, tales como que esas creencias sean (aproximadamente) verdaderas, que tengan fuerza explicativa, etc. *C* es un mecanismo homeostático, un mecanismo uno de cuyos estados-objetivo es que las creencias satisfagan los objetivos cognitivos.

Una de las *funciones* de las preferencias y los deseos es ser racionalmente coherentes; una de las *funciones* de las creencias es satisfacer los objetivos cognitivos. Esto se sigue de nuestra anterior noción de función si realmente son esos mecanismos *P* y *C* mecanismos homeostáticos de este tipo. Supongamos que estos mecanismos homeostáticos producen creencias y deseos investidos de esas funciones. ¿Es *su* función el hacerlo? Eso depende de qué otros mecanismos y procesos producen y mantienen a esos mecanismos formadores de deseos y de creencias. Si esos mecanismos de preferencias y de cognición, *P* y *C*, estuvieran ellos mismos diseñados, producidos o alterados y mantenidos por ingenios homeostáticos cuyos objetivos incluyeran *pretender P* y *C* como ingenios que producen preferencias racionalmente coherentes y creencias aproximadamente verdaderas, entonces tendríamos una doble funcionalidad. Es una función de las preferencias y de las creencias el ser racionalmente coherentes y aproximadamente verdaderas, y también es una función de los mecanismos que las producen el producir cosas así, con esas funciones.

XVIII. Hay un mecanismo homeostático *M*1 cuyo estado-objetivo es que el mecanismo de preferencias *P* dé lugar a preferencias racionalmente coherentes, y *P* es producido y mantenido por *M*1 (a través de la búsqueda, por parte de *M*1, de este estado-objetivo).

XIX. Hay un mecanismo homeostático *M*2 cuyo estado-objetivo es que el mecanismo de creencia *C* dé lugar a creencias que satisfagan objetivos cognitivos, y *C* es producido y mantenido por *M*2 (a través de la búsqueda, por parte de *M*2, de este estado-objetivo).

Resulta plausible pensar que nuestros mecanismos de formación de creencias han experimentado una modelación evolucionaria y social que, parcial pero significativamente, pretendió que tuvieran esas funciones. Hay más. Una vez que la gente adquiere autoconsciencia de sus preferencias y creencias, pueden orientarles, controlar sus desviaciones respecto de la coherencia racional y de la verdad, y proceder a las correcciones oportunas. La percepción consciente se convierte en una *parte* de los procesos *P* y *C*, y los pretende objetivamente al servicio de los objetivos de la coherencia racional y de la verdad.

> **XX.** Uno de los componentes del proceso *P* de formación de preferencias y deseos es la pretensión consciente, por parte de la persona, de llegar a preferencias y deseos racionalmente coherentes.

> **XXI.** Uno de los componentes del proceso *C* de formación de creencias es la pretensión consciente, por parte de la persona, de llegar a creencias que satisfagan objetivos cognitivos.

Esta autoconsciencia, este control, nos confiere una racionalidad más plena. (Algunos podrían sugerir que sólo cuando se satisfacen estas condiciones puede decirse que tenemos racionalidad.)

La percepción autoconsciente puede controlar no sólo preferencias y creencias, sino a los procesos mismos, *P* y *C*, que las forman. Puede modificar y mejorar esos procesos; puede remodelarlos. La percepción consciente se hace así parte de los mecanismos *M*1 y *M*2, y viene así a desempeñar un papel en la determinación de las funciones de los mismos mecanismos de formación de preferencias y de creencias.

> **XXII.** Uno de los componentes del mecanismo homeostático *M*1 que mantiene a *P* es la pretensión consciente, por parte de la persona, de que *P* dé lugar a preferencias racionalmente coherentes.

> **XXIII.** Uno de los componentes del mecanismo homeostático *M*2 que mantiene a *C* es la pretensión consciente, por parte de la persona, de que *C* dé lugar a creencias que satisfagan objetivos cognitivos.

La racionalidad viene, así, a modelar y controlar su propia función. (¿Y viene acaso también la racionalidad autoconsciente a cuestionarse los procesos que producen y mantienen a *M*1 y a *M*2, jugando entonces también un papel en esos procesos?)

Otra posibilidad es una teoría histórica. Un deseo racional sería

un deseo derivado (o que podría serlo) de un conjunto inicial de deseos (biológicamente) dados, y derivado por un cierto tipo de proceso (presumiblemente racional). La racionalidad del deseo sería entonces relativa al punto de partida biológico (y a los procesos de derivación y transformación); los Centauros Alfa podrían haber empezado con un conjunto innato muy distinto de deseos y de reforzadores. ¿Pero debería ser verdad por definición que nuestros deseos biológicamente dados son racionales? Acaso esto ponga un límite para nuestra racionalidad, el hecho de ser *criaturas*. Empezamos con ciertos deseos y predisposiciones; y aunque no estamos para siempre hincados en ellos —podemos modificarlos y transformarlos de varias formas—, siempre nos hallamos en un lugar que resulta accesible partiendo de *allí*.

Permítaseme resaltar de nuevo que mi propósito al discutir estas veintitrés condiciones no es hacer que *éstas* resulten aceptables, ni menos defender sus detalles concretos, sino mostrar qué espacio hay abierto a condiciones de racionalidad de este tipo y mostrar que el cuadro que nos pintó Hume se modifica substancialmente cuando estas condiciones llegan a valer de consuno.

TESTABILIDAD, INTERPRETACIÓN Y CONDICIONALIZACIÓN

Dos rutas alternativas propuestas para dar más contenido a los deseos y a las preferencias racionales me dejan escéptico. La primera se pregunta qué debe ser verdadero si la teoría de la decisión ha de considerarse una teoría empírica con contenido testable, una teoría (de la) que pueda (descubrirse que puede) violar alguien.[23] Cualquier pauta real seguida por las alternativas ya elegidas puede construirse de manera que resulte acorde con las condiciones normativas si se analizan, se dividen y se describen con suficiente detalle los resultados de esa pauta, por ejemplo, «recibir *x* un jueves a las 5 de la tarde de manos de alguien nacido en tal fecha». Si mantener constantes esas descripciones lleva a violar las condiciones, entonces háganse descripciones adicionales detalladas de las alternativas —esto siempre puede hacerse— para resolver la discrepancia. El modo de hacer firme la testabilidad, prosigue el argumento, es introducir una lista exhaustiva de aspectos de las alternativas de los que una persona racional podría pensar que inciden en o afectan a sus elecciones. Nosotros podemos entonces comprobar si sus elecciones satisfacen las condiciones normativas puestas a alternativas fijadas por *esos* aspectos que resulta racional tomar en cuen-

ta. Por lo tanto, si la teoría de la decisión tiene que ser una teoría empírica y testable, presupone preferencias racionales, o al menos, un contenido adicional (racional) para las preferencias.

El argumento es apresurado, según creo. *Cualquier* determinación de las alternativas, de los aspectos que pueden usarse para distinguir alternativas, hará que la teoría de la decisión se convierta en una teoría empírica testable. No es necesario que esa determinación enumere aspectos que haya que tomar *racionalmente* en cuenta. Sin embargo, puede descubrirse que una persona ha violado las condiciones en lo atinente a *estas* alternativas particulares. A decir verdad, (un defensor de) la persona podría reponer que ella *sí* satisface las condiciones normativas en lo atinente a otro conjunto de alternativas. Consideremos ahora la hipótesis existencialmente cuantificada de que hay *algún* conjunto de alternativas, definido para algún conjunto de aspectos, en lo atinente al cual la persona satisface las condiciones normativas. Como mera descripción de sus elecciones pasadas, esto será verdad; alguna que otra definición cuadrará con sus elecciones pasadas. Pero eso no significa que esa definición haya sido la *guía* de sus elecciones. Si la preferencia es, al menos, una disposición a elegir, entonces la tesis de que la persona tiene esas preferencias sobre esas alternativas determinadas compromete a esa persona con las consecuencias ulteriores no sólo de sus elecciones futuras reales, sino también de otras particulares elecciones que *hubiera* podido realizar —suponiendo siempre que sus preferencias no cambien—. La hipótesis tiene consecuencias subjuntivas. No es una verdad trivial, ni una verdad lógica, sino una tesis empírica el que haya algún conjunto de aspectos que definen alternativas de modo tal que una persona realiza y realizaría elecciones acordes con las condiciones normativas impuestas a esas alternativas. Es posible que no exista tal conjunto de aspectos.[24]

¿Cómo habremos, empero, de descubrir si existe o no? ¿Corre la carga de determinar los aspectos particulares, de enunciar la ejemplificación existencial de la hipótesis existencialmente cuantificada, a cuenta de quien niega la tesis de que las particulares elecciones de una persona satisfacen las condiciones, o corre a cuenta de quien la afirma? Dados los costes de negar un enunciado de este tipo cuantificado existencialmente, enormes comparados con la relativa facilidad de mostrar *un* ejemplo cubierto por el enunciado, podría pensarse que la carga de la determinación corre a cuenta de quien sostiene que las condiciones normativas quedan satisfechas, sobre todo si éstas no parecen cumplirse en las determinaciones usuales y obvias. Pero con respecto al argumento que estamos consideran-

do, todo lo que necesitamos observar es que cualquier determinación de las alternativas dará lugar a una hipótesis que es en principio testable. No es necesario determinar las alternativas de una manera que resulte particularmente recomendable a la razón. Considérese la situación a que daría lugar una persona que *no* satisfaciera las condiciones en ninguna alternativa que casara con el listado racional de aspectos propuesto, pero que sin embargo sí satisfaciera esas condiciones en algún *otro* listado (no racional) de aspectos. (Es decir, la persona ha satisfecho las condiciones en *esas* alternativas en sus elecciones pasadas, la determinación ya ha sido ofrecida; inspeccionamos ahora las elecciones futuras de la persona y descubrimos que en realidad continúa satisfaciendo las condiciones en estas alternativas determinadas, que han sido ahora determinadas de un modo que no es una descripción *ad hoc* de datos ya observados, sino una determinación ofrecida de antemano y derivada de la corrección de las predicciones de esa misma determinación.) ¿Estarían realmente dispuestos, los proponentes del argumento que estamos discutiendo, a decir que esta persona no está actuando de acuerdo con la teoría de la decisión sólo por el hecho de que su determinación de alternativas no tiene lo que ellos reputan (quizá atinadamente) como contenido racional?

El segundo argumento en favor de más contenido es un argumento interpretativo. No podemos interpretar a las personas como si ejecutaran acciones intencionales, como si tuvieran algún tipo de preferencias y deseos, a menos que imputemos algún contenido a sus deseos y objetivos, circunscribiéndolos de forma tal que sean de alguna manera como los nuestros, similares a los que nosotros mismos tendríamos en su situación. Tal es la miga, en cualquier caso, de varios y variados principios, propuestos en los últimos años, de interpretación caritativa, aplicada por lo pronto al contenido de los enunciados y de la creencia, y luego, más generalmente, a su combinación con las preferencias.[25] Puesto que estos principios orientativos de interpretación no me parecen adecuados ni siquiera en su aplicación a los enunciados y a las creencias, aprovecharé la ocasión de discutirlos aquí, en éste que es su contexto más robusto, más que en su aplicación a las preferencias y a los deseos. (Vale la pena discutir este punto de vista con algún detenimiento, aunque sólo sea por la habitual prevalencia de que goza aplicado a otros tópicos filosóficos.)

La raíz de mi insatisfacción es que, al conceder un peso indebido a las posiciones que ocupamos, a nuestras propias creencias y preferencias, estos puntos de vista parecen imperialistas. Se podría

tratar de evitar eso suponiendo que todos compartimos un trasfondo evolucionario común. Eso limitaría el alcance general de la tesis, la cual a partir de entonces se propondría como una tesis acerca de (las creencias y las preferencias de) todos los agentes racionales, en la tierra o en algún otro lugar del universo. Puesto que no hay ningún deseo aislado que sea necesario para servir a la adaptación inclusiva en todos los medios, no parece éste entonces un nivel de generalización que resulte iluminador, no se puede predecir a partir de él que *todos* los organismos evolucionados compartirán *algunos* deseos. Si nuestras propias creencias o preferencias carecen de autoridad por medio de este argumento interpretativo, que las convierte en hitos para todas las creencias y todos los deseos. Hasta aquí las razones generales del escepticismo; vayamos ahora a los detalles.

Al interpretar (o traducir) lo que alguien dice o escribe no funcionaría hacerlo lo más *verdadero* posible, pues podemos saber que esta persona no estaba en una posición que le hubiera permitido aprender lo que hemos aprendido nosotros acerca de la verdad; no había realizado experimentos, colectado datos, etc. Más razonable sería

> 1. Interpretar o traducir lo que la persona dice y hace de tal modo que la persona parezca lo más *racional* posible.[26]

Esta propuesta entraña el supuesto de que la persona usa los mismos criterios de racionalidad que nosotros. Sin embargo, hay una robusta evidencia *prima facie* de que algunas personas no razonan o sacan inferencias de acuerdo con (lo que nosotros consideramos) criterios correctos.[27] También hay evidencia de que ciertas drogas hacen que los procesos de formación de creencias de las personas sean menos racionales. Y algunas personas declaran explícitamente que siguen principios de racionalidad distintos.[28] Si les imputáramos nuestros criterios de racionalidad, parecería que estamos distorsionando sus puntos de vista. Además, los criterios de racionalidad de una sociedad pueden cambiar y desarrollarse con el tiempo. Se ha dicho que la alfabetización trae consigo una acrecida sensibilidad respecto de la coherencia y del apoyo razonado —los enunciados escritos se pueden examinar repetidamente y comparar con otros—, y que por lo mismo, lleva a la formulación de nuevos criterios de razonamiento y de crítica.[29] Un antropólogo podría, en consecuencia, dudar de imputar ciertos criterios de creencia racional a las sociedades sin escritura e interpretar a sus miembros como si satisfacieran esos criterios.

Una posición en repliegue podría sostener que los adultos pueden ejemplificar y exhibir distintos criterios de racionalidad como resultado de haberles sido inculcados culturalmente y de sus propias experiencias vitales, pero que todos los niños parten de los mismos modos iniciales de formación de creencias, de las mismas capacidades y predisposiciones para formar creencias. Todas las creencias de los adultos son el resultado racional de esos procesos iniciales de formación de creencias y deseos, según operan en las experiencias y en los datos que la vida individual y la cultura ponen en su camino. Aun cuando no hay diferencias iniciales entre individuos en lo que hace a sus disposiciones a formar creencias o a adquirir nuevas disposiciones, sus diferencias en las experiencias vitales pueden llevarles no sólo a diferencias ulteriores en determinadas creencias y en determinados deseos, sino también a diferencias en sus modos y procedimiento de adquisición de creencias y deseos y en sus formas de actuar sobre esas creencias y esos deseos. Y algunos de estos procedimientos diferentes podrían estar *inculcados* por las diferentes experiencias, no simplemente inferidos de ellas de acuerdo con los procedimientos previos. Una teoría científica podría explicar todas esas diferencias resultantes (y cualesquiera diferencias iniciales que pudiera haber), pero ello no significaría decir que se podría aplicar alguna teoría de la racionalidad en cada etapa o en alguna de ellas. (El término *racional* podría no tener aplicación a la inicial etapa común infantil de ir amontonando creencias.)

Donald Davidson ha sostenido que no podemos dar un sentido coherente a la noción de esquemas conceptuales radicalmente diferentes. Su criterio de diferencia radical es el sorprendente criterio de no-traducibilidad —no al modo en que yo normalmente usaría la noción de «esquema conceptual diferente»—, pero sin embargo el argumento parece implicar que alguna traza de racionalidad debe funcionar en cualquier cultura linguística.[30] No obstante, Davidson mismo proporciona un contraejemplo a su propia concepción: es posible que la traducibilidad entre esquemas conceptuales no sea transitiva. Podríamos ser capaces de traducir a los saturnianos, que son capaces de traducir a los plutonianos, y sin embargo no ser capaces de traducir a los plutonianos, cuyo esquema conceptual nos resulta ininteligible. El intento que hace Davidson de potenciar este ejemplo es débil: se pregunta cómo es posible que sepamos que los saturnianos *traducen* del plutonés. He aquí cómo. Están haciendo lo mismo, es decir, usando el mismo procedimiento (y meta-procedimiento) que usan cuando traducen del inglés al saturnés, y este procedimiento consiste en proyectar cada sentencia *Ei* del inglés en la

misma sentencia *Si* del saturnés que *nosotros* proyectamos (cuando traducimos) en *Ei*. Tenemos tanta razón para pensar que están traduciendo como para pensarlo cuando nosotros lo estamos haciendo.

Sin embargo, ¿acaso no podemos traducir del plutonés traduciendo al inglés las sentencias saturnesas que los saturnianos han traducido del plutonés? En tal caso, ¿no entenderíamos a los plutonianos, después de todo? Y podría argüirse que, aun si los mediadores saturnianos no existieran realmente, la posibilidad de su existencia basta para mostrar que es posible para nosotros entender y traducir inteligiblemente a los plutonianos, que es todo lo que la tesis davidsoniana necesita.[31] La noción «directamente traducible por» es no transitiva. Pero el que dos grupos o dos individuos, no directamente traducibles entre sí, se hallen en una cadena de traducciones directas a través de intermediarios, ¿basta para constituir una traducción indirecta y, por lo tanto, una traducción *simpliciter*? Yo creo que no.

La razón de que la relación «directamente traducible por» sea no transitiva es que (sólo) requiere un solapamiento grande, pero no exacto, entre las condiciones de aplicación de los conceptos o enunciados, y una cadena de solapamientos suficientes puede llevar a un solapamiento insuficiente entre sus puntos extremos. De aquí que podamos imaginar dos cadenas intermedias distintas de traducción directa a través de diferentes civilizaciones interplanetarias; cada eslabón de esas cadenas exhibe un solapamiento suficiente. A través de una cadena de transformaciones graduales de gran solapamiento que empiezan en la fuente original *S*, traducimos el enunciado *y* de nuestro vecino de cadena más cercano, *Y*, como *z*; a través de otra cadena de transformaciones graduales de gran solapamiento que empieza en la misma fuente original *S*, traducimos el enunciado *p* de nuestro vecino *X* como *q*. Sin embargo, *z* y *q* pueden ser tan diferentes como ustedes quieran, tan diferentes como nuestros conceptos «vaca» y «entropía».* Por lo tanto, se puede demostrar la

* He aquí una analogía. En un sistema de procesamiento paralelamente distribuido, algún espacio que represente a los vectores activados de las unidades intermedias como subregiones puede tener una secuencia de regiones que se solapan, estando cada región lo suficientemente cerca de otra para actuar como un atractor sobre algún que otro vector en la otra región; sin embargo, la primera y la última pueden ser bastante diferentes y distintas, no estar lo suficientemente cerca como para atraer un vector en la región que se halla al otro extremo de la cadena. Dos cadenas de este tipo, que van en direcciones diferentes, llevarían, como destino final, a dos regiones muy diferentes. La región en la que empiezan las dos cadenas diferentes quizá no sea describible conceptualmente en nuestro propio vocabulario (que es distinto de un vocabulario que habla explícitamente de pesos y de vectores de activación), aunque las dos regiones en que desembocan las cadenas sí tengan descripciones conceptuales que difieren enormemente entre sí.

mutua ininteligibilidad imaginando la existencia de *algunas* cadenas intermedias de traducciones. Hay que probar que no puede haber dos cadenas de este tipo que sean lo suficientemente diferentes como para llegar a traducciones radicalmente diferentes. Pero es claramente el caso que este tipo de cadenas distintas pueden existir; y si supiéramos de ellas, ¿acaso no diríamos que no podemos entender en absoluto a la civilización en este punto final?*

También podríamos tener modos de catapultarnos a nosotros mismos hasta esquemas conceptuales radicalmente diferentes, los cuales podríamos entonces usar pero no traducir; catapultarnos, por ejemplo, hasta esquemas místicos merced a ciertas drogas u otros procedimientos para generar experiencias místicas. Alasdair MacIntyre sugiere que podemos ir a otra sociedad y aprender sus modos y su lenguaje desde el comienzo, lo mismo que un niño, para luego descubrir que no podemos traducir su lenguaje al inglés.[32] Podríamos entonces ser capaces de usar otros esquemas conceptuales que, de acuerdo con nuestros criterios, o con cualesquiera otros que fuéramos capaces de formular, no son racionales.

Ronald Dworkin ha propuesto otro principio orientativo de interpretación.

> **2.** Interpreta (y traduce) de manera que lo que la persona diga y haga parezca lo mejor, lo más justificado, guiado por principios y correcto posible.[33]

Dworkin propone esto primordialmente como principio de interpretación de las reglas en un contexto institucional, y cree que resulta plausible también como principio de interpretación literario. En ese modo de interpretación, las cosas se ven «bajo su mejor luz», bajo la mejor luz posible. Podríamos llamar a este principio de dar el mayor lustre a todo el Principio Panglosiano. Dworkin formularía el principio de tal manera que no pudiera torcerse en un sentido compasivo con las instituciones genocidas o esclavistas. Mas ¿no puede llevar a confusión también en el caso de instituciones (de las que sabemos) que representan *compromisos* entre intereses competiti-

* Quizá podríamos tener alguna razón para pensar que una de estas largas cadenas es más intrincada y zigzagueante que la otra. Supongamos, pues, que no es éste el caso; en el dilatado espacio conceptual, sea cada senda igualmente directa. La discusión podría continuar preguntando cómo conocemos que cada eslabón a lo largo de la cadena, aun el más distante, tiene que ver con una *traducción*. ¿Debe éste traducir directamente de un modo suficientemente cercano al nuestro, o basta con que traduzca indirectamente?

vos, o que fueron modeladas por los resultados *velis nolis* de ensayos de tomar o mantener el poder? ¿No entenderíamos así mejor una institución como el resultado de los procesos que realmente la produjeron, no a través del prisma de un intento de racionalización?

Examinemos otro principio orientativo:

> **3.** Interpreta (y traduce) de manera que lo que la persona diga parezca lo más inteligible posible.

Esto da margen también para los otros principios orientativos, pues la verdad, la racionalidad y la bondad son *modos* de ser inteligible. Pero no son los únicos modos, y también pueden resultar inteligibles las desviaciones respecto de esos modos. Pero no siempre queremos traducir de manera que el contenido de lo dicho parezca lo más inteligible posible. La persona podría haber tomado una droga, y nosotros, saber que esa droga tiende a producir sinsentidos ininteligibles en el habla de la gente.

Hay que modificar, pues, el principio 3:

> **4.** Traduce «p» de manera que resulte lo más inteligible posible el hecho de que en este contexto esta persona diga «p».

Éste es el hecho global que hay que hacer inteligible, el hecho de que en este contexto esta persona diga «p». El principio 4 ofrece una orientación sobre el modo de traducir «p», y deja también un margen para incluir lo que hacen los historiadores y los antropólogos cuando consideran el contexto histórico y cultural, así como los propósitos y los motivos de las personas, sus temores y sus supersticiones.

Podemos ampliar el alcance. La persona no sólo dice «p», sino también «q», «r» y «s». ¿No deberíamos traducir de manera que resultara lo más inteligible posible el hecho de que en esos varios contextos la persona dijera todas esas cosas particulares? Sus contemporáneos y sus antecesores dijeron otras cosas también. ¿No deberíamos traducir de manera que resultaran lo más inteligibles posible los hechos de todos sus decires? ¿Podría esto trastornar o modificar a aquello que haría máximamente inteligible el hecho (aisladamente considerado) de que la persona diga «p»?

Los anteriores principios —traducir de manera que maximice la verdad, o la racionalidad, o la bondad— eran estimulantes, pero inadecuados. El principio 4, por otro lado, parece aburridamente obvio y poco iluminador. «*Evidentemente,* tenemos que traducir e in-

terpretar de manera que lo que se dice resulte lo más inteligible posible».

¿A quién debemos hacer lo más inteligible posible el hecho de que la persona diga esto? A nosotros mismos, traductores e intérpretes. Por lo tanto, nuestros criterios de inteligibilidad entrarán —junto con nuestras teorías del habla, de la acción, de la vida y de la sociedad— en el juicio que hagamos de cuál sea la interpretación que torne más inteligible el hecho. Sin embargo, los puntos de vista de la persona no *importan* nuestros criterios, ni podemos *atribuírselos* —aunque los principios que hacen que sus puntos de vista parezcan lo más racionales o buenos posible habrían hecho precisamente esto—. El hecho de «hablar en jerga» puede resultar inteligible sin que lo resulte aquello de que se habla. Resulta inteligible el decir, no (necesariamente) lo dicho.

Un modo de hacer tan inteligible como sea posible el hecho de decir lo que dijo la persona es *explicar* lo mejor posible su decirlo. Esa explicación se fundará en nuestras teorías de la conducta humana y en nuestro sentir común acerca de la misma, y hará que su habla case con ambos. Cuanto más detallada nuestra red explicativa, cuantos más detalles de su hablar y de su habla quepan en esa red, tanto más inteligibles habremos hecho a estos últimos. Mostrar su propósito al decir lo que dijo, identificar qué acto ilocucionario ejecutó al decirlo: todo ello puede contribuir a la inteligibilidad de su decirlo.

¿Podemos también ayudar a hacer lo más inteligible posible el hecho de su decir esto mostrando que lo que dijo es lo más verdadero, o racional, o bueno posible, o dicho en otras palabras, secundando los principios orientativos previamente discutidos? Sólo si tenemos razones para pensar que, precisamente en aquel momento, la persona estaba siendo cuidadosa, o racional, o buena.* Si tenemos

* Hace algunos años, criticando el modelo de explicación nomológica (deductiva o estadística) de C.G. Hempel como condición necesaria para la explicación histórica, William Dray sostuvo que había otro modo de explicación histórica, la explicación racional, que no usaba leyes, sino que explicaba una acción mostrando qué era lo (o algo) que había que hacer racionalmente en esas circunstancias. (Véase William Dray, *Laws and Explanation in History* [Londres: Oxford Univ. Press, 1957].) Hempel replicó que en este tipo de explicación necesitábamos añadir que el agente tenía la disposición a ser racional y la ejercía, es decir, que se trataba de una situación en la que él *haría* lo racional. Por lo tanto, las explicaciones de Dray implican una ley implícita; para un agente y para una gama de circunstancias o situaciones del tipo *S*, el agente haría lo racional en una situación de tipo *S*. Y la explicación continuaría; la persona estaba en la situación *si*; *si* era una situación de tipo *S*; en *si*, a acción *A* era la opción racional; por lo tanto, hizo *A*. (Véase C.G. Hempel, «Ra-

razones para pensar que *no*, entenderlo como cuidadoso, o racional, o bueno generará desconcierto, no inteligibilidad.

Supongamos, sin embargo, que alguien propone que lo que haya que hacer lo más inteligible posible *no* sea el hecho del decir de la persona, sino *lo que ésta dice*. ¿Puede esto convertirse en otro objetivo de la interpretación, e implicará hacer al enunciado verdadero, correcto, justificado o racional? Sin embargo, las falsedades, las malas directrices y los insultos son todos bastante inteligibles —es decir, su *contenido* es completamente inteligible—. Lo que podemos no entender es por qué alguien dice estas cosas o las proclama, pero esto nos lleva de regreso a la explicación de que fueran dichas.

Con todo, ¿acaso no entendemos algo mejor cuando vemos la mejor justificación que podría dársele —pueda o no ofrecerla la persona que lo dijo—, cuando vemos lo que *podría* decirse en favor de ello? Yo no negaría esto; también entendemos mejor un enunciado cuando vemos lo que podría decirse en su contra. Al descubrir sus puntos fuertes y sus puntos débiles, al abarcar sus conexiones en otras teorías y con otros problemas, exploramos más concienzudamente su naturaleza. La interpretación más caritativa presenta sólo una cara de la naturaleza de algo.

Obsérvese la analogía entre las cuestiones que hemos discutido y la reciente discusión, en teoría evolucionaria, entre quienes pretenden ofrecer explicaciones adaptativas para una amplia gama de rasgos —que habría que entender como rasgos seleccionados *positivamente*— y quienes pretenden explicar muchos rasgos sin necesidad de imputarles valor adaptativo alguno, cargándolos en el haber de otras vías evolucionarias, por ejemplo, entendiéndolos como efectos laterales de otros rasgos que *fueron* seleccionados positivamente, o como productos de la necesidad de satisfacer varias restricciones estructurales, o como resultado de la deriva genética, etc. Pensar que el *único* modo de entender un producto intelectual es interpretarlo como el más racional o bueno posible es análogo a pensar que el único modo de explicar un rasgo de un organismo es mostrar que es óptima o máximamente adaptativo. Nuestro principio

tional Action», *Proceedings and Addresses of the American Philosophical Association* 35 [1961-1962]: 5-23.) Para *algún* individuo, lo que podría requerir más explicación es por qué actuó racionalmente esta vez. Puesto que casi siempre actúa irracionalmente, no basta con decir simplemente que la acción que realizó era la acción racional. ¿La realizó accidentalmente? ¿Tiene tendencia a actuar racionalmente en una gama reducida de situaciones y era ésta una de ellas? ¿Tuvo un ataque momentáneo de racionalidad?

de interpretar la generación de este producto intelectual lo más inteligiblemente posible se corresponde con la idea que reconoce muchos modos diferentes de explicación evolucionaria y considera una cuestión empírica abierta cuál de ellos será de aplicación en un (tipo de) caso dado.[34] Bien que las cuestiones son estructuralmente paralelas, aún podrían sin embargo perseguirse diferentes estrategias explicativas en los dos ámbitos, acaso porque se crea que las presiones selectivas son más fuertes en el reino biológico.

Permítaseme mencionar otra posible ruta para desarrollar condiciones sobre el contenido de las preferencias. Se han propuesto condiciones bayesianas para la revisión de las probabilidades en el transcurso del tiempo, y bien podría tratar de emularse a éstas imponiendo a la utilidad condiciones intertemporales. Así como la probabilidad condicional, fundada en la evidencia *e*, que la hipótesis *h* tenía ayer se supone que arroja hoy, cuando la evidencia *e* acaba conociéndose, una nueva probabilidad para *h* que es igual a su probabilidad condicional anterior, así también podríamos desarrollar una noción de utilidad condicional y proponer una condición análoga de condicionalización intertemporal.[35]

Sin embargo, si las particulares probabilidades o preferencias de ayer no tienen ninguna autoridad especial —todo lo que se les exige que hagan es satisfacer los axiomas de la teoría de la probabilidad, o las condiciones estructurales impuestas a las preferencias por Von Neumann-Morgenstern—, entonces ¿por qué habrían ellas de ejercer autoridad alguna sobre las particulares probabilidades o preferencias de hoy? Las condiciones estructurales impuestas a las preferencias en cualquier momento dado dicen que las preferencias deben ir de consuno de una cierta forma; si no van, se deja abierta la cuestión de cuáles de ellas deben sufrir revisiones. De la condición normativa de que las preferencias sean transitivas, junto con las premisas de que una persona prefiere *x* a *y* y prefiere *x* a *z*, *no se puede* derivar la conclusión de que la persona debería preferir *x* a *z*. Quizá no debería preferir *x* a *y*, o *y* a *z*. La exigencia de que las preferencias sean transitivas *no* debería interpretarse en el sentido de que si una persona prefiere *x* a *y* y prefiere *y* a *z*, entonces esa persona debería preferir *x* a *z*. La exigencia de transitividad debería ser, en cambio, interpretada así: no debería darse el caso de que una persona prefiera *x* a *y*, prefiera *y* a *z*, y no prefiera *x* a *z*. De esa conclusión no puede derivarse ninguna conclusión por separado acerca de qué particular preferencia debería tener una persona.[36]

Considérese la siguiente versión intertemporal (putativa) de la

condición estructural de que las preferencias sean transitivas: si el lunes una persona prefiere *x* a *y* y prefiere *y* a *z*, y no tiene preferencia entre *x* y *z*, entonces el martes esta persona debería preferir *x* a *z*. Ésta sería una condición objetable. Quizás el martes la persona debería más bien dejar de preferir *x* a *y*, o *y* a *z*. Evidentemente, el martes la persona no puede cambiar el hecho de que el lunes prefiriera *x* a *y* e *y* a *z*. ¿Se puede argumentar como sigue?: aunque es verdad que la condición, correctamente interpretada, sólo dice en *un* momento dado que las preferencias no deberían ser discordantes de una determinada manera, y deja abierta la cuestión de cuáles de ellas deberían ser objeto de revisión para conseguir la concordancia; sin embargo, una condición intertemporal, combinada con el hecho de que el pasado es inalterable, arroja ahora la conclusión de que la persona *debería* tener hoy una determinada preferencia. También este argumento sería objetable. Supongamos que el martes la persona prefiera *z* a *y* y desea no haber preferido el lunes *x* a *y*, o *y* a *z*. Puesto que no hay nada que pueda hacer para alterar esos hechos pasados, ¿significa eso que debe ahora instaurar un hecho presente preferir *x* a *z*, al que también pondría reparos? Sin duda, no.

No obstante, tenemos el conocido argumento del bombeo de dinero: una persona con preferencias intransitivas que está dispuesta a realizar pequeños pagos para lograr su alternativa más preferida puede ser inducida, mediante una secuencia de esos pagos, a completar el círculo, acabando donde empezó, pero más pobre. (Y el ciclo puede repetirse.) Intertemporalmente también; la persona que el lunes prefiere *x* a *y* e *y* a *z*, pero que el martes prefiere *z* a *x*, si empieza con *z* el lunes, puede verse inducida a pagar dinero para acabar donde empezó. (El lunes paga para moverse de *z* a *y*, y luego paga otra vez para moverse de *y* a *x*, y el martes paga para moverse de *x* a *z*.) Sin embargo, este hecho no puede demostrar que la persona debería preferir *x* a *z* el martes, o que debería actuar como si lo prefiriera; pues no lo prefiere, ni necesita hacerlo. Sus preferencias pueden *cambiar*, no meramente desarrollarse como implicaciones de sus preferencias previas. Sería absurdo imponer una condición normativa intertemporal que prohibiera a alguien invertir una preferencia, aun en el caso de que quien lo hiciera acabara efectuando pagos en los días siguientes para acabar donde empezó. Tal condición normativa no se hace más plausible por el hecho de ser un análogo intertemporal de la condición de que la preferencia sea asimétrica.

Objeciones similares pueden hacerse al uso de la noción de uti-

lidad condicional en una exigencia intertemporal adicional de condicionalización —nadie la ha propuesto realmente, después de todo—, así como al reputado criterio, según el cual las probabilidades deberían evolucionar en el tiempo de acuerdo con el principio de condicionalización bayesiana.[37] De acuerdo con la doctrina de los bayesianos estrictos, para quienes la única noción admisible de probabilidad es la que mide el grado de creencia de una persona, esas probabilidades personales sólo tienen que satisfacer los axiomas de la teoría de la probabilidad. Cualesquiera imputaciones particulares de probabilidad, cualquier combinación de grados de creencia que los satisfagan serán igualmente buenas.[38] Por lo tanto, las particulares probabilidades del lunes no tendrán especial autoridad. El bayesiano tiene un argumento para explicar por qué las probabilidades en cualquier momento dado deberían satisfacer los axiomas de la teoría de la probabilidad. Se trata del argumento del libro holandés para las probabilidades, que se corresponde con el argumento del bombeo de dinero para las preferencias; a una persona cuyas probabilidades, cuyos grados de creencia, no satisfacen los axiomas de la teoría de la probabilidad, y que está dispuesta siempre a apostar de acuerdo con sus grados de creencia, se le puede ofrecer una serie de apuestas que ella estaría dispuesta a aceptar, tales que, de los resultados de todas las apuestas tomadas conjuntamente, no puede sacar ningún dinero y (según qué axiomas anden en juego), o bien puede perder algún dinero, o bien perderá sin duda algún dinero.*

También hay una versión intertemporal de este argumento del libro holandés; a una persona cuyas probabilidades no evolucionan a lo largo del tiempo de acuerdo con el principio de condicionalización se le puede ofrecer una secuencia de apuestas, a través de la cual la persona no puede ganar ningún dinero y debe acabar perdiéndolo.[39] Pero consideremos el caso de una persona que decida, por cualquier razón, mantener las probabilidades de hoy que no están en la deseable relación condicional con sus probabilidades de

* El hecho *F*1 de que se *ofrezca* una apuesta sobre *p*, cuando no es en general verdadero que se ofrezca una apuesta sobre toda proposición *q*, puede alterar la probabilidad personal de *p* en relación con la que tenía anteriormente. La conducta apostadora es ahora indicio de la probabilidad personal imputada a *p*, dado el hecho *F*1. Supongamos que la apuesta no la ofrece otra persona, sino que existe una situación con una acción *A* accesible cuyas ganancias se correspondan con las de alguna apuesta sobre la verdad de *p*. Lo que indica la realización de esta acción es la probabilidad personal de *p* dado el hecho *F*2, en donde *F*2 es el hecho de que tal acción «apostadora» es accesible respecto a *p* cuando no es accesible respecto a cualquier otra proposición *q*. ¿Bajo qué condiciones podemos obtener la probabilidad tal como era antes, sin contaminarla ni alterarla?

ayer. Esa persona puede desear que sus probabilidades dependan unas de otras a lo largo del tiempo, de manera que puede desear que sus probabilidades de ayer hubieran sido distintas, pero desgraciadamente es demasiado tarde para cambiarlas. ¿Debería ahora estar ligada por el peso muerto del pasado, por probabilidades que, después de todo, quizá no tengan en su favor sino el hecho de que, conjuntamente, satisfacieron los axiomas (atemporales) de la teoría de la probabilidad? Mas esto significaría que la persona tendría hoy un grado de creencia que no desea tener y apostaría hoy a juegos en los que no desea apostar. (¿Debería tratar las probabilidades de ayer como un coste sumergido?) Sin duda, la persona puede actuar de acuerdo con las probabilidades de hoy y sentirse afortunada de que nadie le ofreciera ayer ciertas apuestas. Aun si hubiera realizado ayer ciertas apuestas, eso no significaría que hoy estuviera obligada a apostar fundándose en las implicaciones de las creencias de ayer que ahora rechaza, trocando buena por mala moneda. (Supongamos que la persona piense que las probabilidades de ayer fueron el resultado de un estado febril, o de la creencia —engañosa, según la ve ahora— de que Dios le estaba hablando directamente. ¿Mantendría el bayesiano que esta persona debe condicionalizar hoy sobre aquellas probabilidades? ¿O diría acaso el bayesiano que las de ayer no estaban condicionalizadas sobre las de anteayer, de manera que hoy podemos ignorarlas? El martes, la persona no condicionaliza sobre sus probabilidades del lunes; cuando llega el miércoles, ¿tiene que condicionalizar sobre esas probabilidades mal derivadas del martes, o debería condicionalizar sobre las del lunes aprovechando la nueva evidencia que ha ganado desde entonces?) La persona puede también rechazar sus probabilidades anteriores porque entretanto ha concebido hipótesis nuevas que son alternativas a una hipótesis dada; el dar más estructura a la categoría de «alternativa», y el ver cuál es el contenido detallado de algunas alternativas, le lleva a reconceptualizar un ámbito, y así, a reasignar las probabilidades de un modo que no significa condicionalización de las anteriores basada en información nueva, sino que se parece más a una nueva asignación de probabilidades iniciales.[40]

Aun si la persona realizó ayer realmente apuestas basadas en sus probabilidades de aquel momento, podría, sin embargo, querer apostar basándose en sus juicios de probabilidad de hoy. Es verdad que esto le garantizará la pérdida de *algún* dinero, pero puede preferir una pequeña pérdida cierta a una gran pérdida probable, si esto último es lo que va a ocurrir, de acuerdo con su concepción de hoy, en caso de apostar hoy fundándose en probabilidades que ahora con-

sidera incorrectas, las probabilidades basadas en la condicionalización de las probabilidades de ayer.

Mas nuestra posición respecto de las preferencias no es la posición de los bayesianos estrictos respecto de las probabilidades personales; éstos no imponen más condición a las probabilidades de cualquier momento dado que la satisfacción de los axiomas del cálculo de probabilidades. En la anterior sección, propuse para las preferencias condiciones que iban más allá que las condiciones normativas estructurales habituales. Si las preferencias y las utilidades de la persona satisfacían ayer estas condiciones adicionales, ¿no bastaría eso para bloquear la observación, según la cual «quizá la persona debería haber tenido preferencias distintas ayer», y para justificar así la condicionalización basada en las preferencias y utilidades que la persona realmente tenía entonces? Pero también hay otras preferencias que habrían satisfecho esas condiciones adicionales además de las estructurales. Todo lo que parece necesitarse es que la persona llegue hoy a algún conjunto de preferencias satisfactorias, esto es, a preferencias que satisfagan todas las condiciones impuestas y que hubieran podido surgir también por la vía de la condicionalización a partir de algún que otro conjunto de preferencias satisfactorias de ayer. Pero no es necesario que este último conjunto sea el particular conjunto que la persona tenía realmente ayer. Y no parece que esta debilísima condición intertemporal imponga ningún contenido adicional al anterior requisito de que las preferencias de una persona sean satisfactorias en un momento dado, esto es, que satisfagan en el presente las condiciones estructurales y adicionales propuestas antes para las preferencias.

Ninguna de las avenidas exploradas en esta sección ha llegado, en mi opinión, a establecer condiciones adicionales adecuadas para el contenido de las preferencias. Yo creo que deberíamos ir más allá del restricto punto de vista de Hume, y en la sección anterior he propuesto, muy tentativamente, algunas condiciones de contenido. Para mí sería bienvenida una teoría de la racionalidad más potente respecto del contenido de las preferencias y los deseos, una teoría que constriñera y restringiera más lo dicho aquí, aunque no consiguiera fijarlo completamente.

Kant trató de derivar objetivos de la noción misma de racionalidad, entendida a su modo como adhesión a principios. Ya dijimos antes que los principios son ingenios con funciones definidas, diseñados para trabajar en equipo con (y modificar) objetivos que nos vienen dados. El intento kantiano de divorciar los principios de *cualesquiera* objetivos dados, para luego derivar objetivos, o determi-

nadas restricciones a las máximas de conducta, de la desnuda noción de adhesión a un principio, fracasa. Si nuestros procedimientos racionales fueron diseñados para trabajar en equipo con objetivos biológicamente dados (o con deseos y objetivos, la satisfacción de los cuales estuviera correlacionada en nuestro pasado evolucionario con el robustecimiento de la adaptación inclusiva), entonces no resulta sorprendente que fracase un intento de derivar racionalmente objetivos *de novo,* sin partir de *ningún* objetivo o deseo. Eso no significa que estemos hincados en los deseos y en los objetivos con los que empezamos. Nuestros procedimientos racionales nos permiten modificarlos significativamente, dando pasos que, considerados uno a uno, son pequeños, pero que, repetidos, se acumulan hasta desembocar en un gran cambio.

Supongamos que alguien tuviera éxito en la tarea de alumbrar una teoría plenamente adecuada de la racionalidad substantiva de los deseos. Dada esa teoría, nos encontraríamos en situación de decir que una acción o un procedimiento para formar creencias es racional cuando es *instrumentalmente* efectivo en punto a lograr (no cualquier objetivo arbitrario, sino) objetivos *racionales.* Es verdad que eso sería una gran mejora. La racionalidad instrumental ya no agotaría el entero dominio de la racionalidad; podríamos contraponerle la racionalidad substantiva de los objetivos. Mas, aun con esa modificación, la racionalidad seguiría siendo en gran medida instrumental.[41] Cuando decisivamente rebasamos esta amplia estructura instrumental es cuando damos el paso hacia el valor decisional.

HEURÍSTICA FILOSÓFICA

Lo que hacen los científicos y los filósofos cuando construyen teorías es pensamiento racional: formulan problemas intelectuales, piensan en posibles soluciones para esos problemas, someten a test y evalúan esas soluciones, etc. Este pensamiento puede ser instrumental, pero no está primordialmente orientado a la creencia, ni siquiera a la creencia sobre cuál sea la mejor solución del problema. De manera que el tipo de razones a favor y en contra que antes discutimos no están aquí en primera fila. Sin embargo, una teoría de la racionalidad debería ser capaz de cubrir e iluminar este tipo de actividad teórica.

Los filósofos que se ocupan del razonamiento tienden a concentrarse en una gama exageradamente estrecha de pensamiento, y a considerarla el único modo legítimo de razonar. (Y a veces, los espe-

cialistas en racionalidad consideran a ésta como un mero mecanismo exclusionario, cuyo principal propósito es descartar algo como «irracional», en vez de tratarla positivamente como un vehículo eficiente y efectivo.) Si el mejor razonamiento que los filósofos y los científicos despliegan cuando crean sus teorías no consigue encajar con esa estrecha área que se ha considerado legítima, entonces está claro que esa caracterización del pensamiento racional anda necesitada de expansión.

Pensar una nueva teoría podría entenderse como solventar un problema intelectual, acaso inmediatamente después de percibirlo o descubrirlo. Hay dos áreas interrelacionadas que investigar, pues: plantear el problema y resolverlo. El estudio del planteamiento de problemas explora el modo en que llegan a descubrirse y a formularse los problemas fértiles, qué rasgos caracterizan a una situación de problemas y qué factores los desencadenan o los moldean. El estudio de la solución de los problemas atiende a lo que una persona hace y logra *dentro* de esta situación de problemas: la pauta seguida por sus intentos de solventar el problema o de transformarlo, el progreso y los escollos, el modo en que a ella le parece que el producto final avanza o se quede corto con respecto a sus objetivos. (También podemos estudiar la medida en que la persona avanza en o resuelve el problema importando otros criterios y conocimientos distintos de los suyos.)

¿Qué es un problema intelectual? La literatura de la inteligencia artificial y de la heurística formal ofrece una útil caracterización de la estructura general de un problema. Un *problema bien definido* es un problema para el que están explícitamente determinados y delimitados cada uno de los siguientes rasgos.

1. Un *objetivo*, un criterio evaluativo para juzgar resultados y estados.
2. Un *estado inicial*, que consiste en una situación (de partida) y en los recursos de que disponemos.
3. *Operaciones admisibles*, que pueden usarse para transformar estados y recursos. Esas operaciones admisibles se formulan en términos de reglas que pueden aplicarse para transformar el estado inicial y luego seguir transformando una y otra vez los estados transformados resultantes.
4. *Restricciones* en relación con los estados intermedios por los que puede pasarse mientras se hace camino, en relación con los estados que pueden alcanzarse, en relación con las operaciones que pueden hacerse, cuántas veces, en qué orden, etc.
5. Un *resultado*, un estado final.

Una *solución* al problema es una secuencia de operaciones admisibles que transforma el estado inicial en un resultado que satisface el objetivo sin haber violado ninguna restricción en algún momento a lo largo del camino.[42] Ejemplos de problemas bien definidos son rompecabezas formales tales como el de los «misioneros y los caníbales» —ya podríamos hoy encontrar un nombre mejor— y tareas de prueba formal («empezando con estos axiomas y usando sólo estas reglas de inferencia, pruebe este teorema»).

Esta lista de componentes resulta iluminadora, aunque la mayoría de los problemas con que se enfrenta la gente no pueden determinarse con tanta exactitud.* A menudo, la gente no *se enfrenta* sim-

* En el modelo de problema, un objetivo es un criterio evaluativo, tal que, dado un resultado cualquiera, hay una respuesta clara, determinada y viable para cuestión de si este resultado satisface el criterio. Sin embargo, en la vida real puede no estar claro si un resultado satisface un objetivo, y hasta qué punto. El objetivo puede ser un objetivo a largo plazo, y el resultado más inmediato, hacer que el logro del objetivo sea más o menos probable. Una evaluación puede ser comparativa y estar en manos de otros, como cuando los planos de un arquitecto se presentan a un concurso. Por lo tanto, las evaluaciones subrogadas, las evaluaciones provisionales, pueden usarse como estimaciones de una evaluación final. También puede haber más de un objetivo relevante, y acaso una ordenación ponderada o jerarquía entre ellos, pero a menudo las prioridades serán más vagas; los objetivos no vendrán ordenados, y la lista misma de ellos puede estar abierta. (Al desarrollar un proyecto, alguien podría darse cuenta de que se toma una dirección que efectivamente desemboca en cierto objetivo que no le llamó la atención previamente y que, en cambio, ahora, *se convertirá* en un objetivo o propósito del trabajo.) También es posible que los objetivos no estén predeterminados; una persona puede seleccionar sus objetivos, o modificar los que le venían dados.

Los recursos pueden estar disponibles en diferentes grados, ser algunos más difíciles de conseguir y de usar que otros, tener algunos más costes que otros. Es posible que la situación de partida de una persona no le resulte clara. También es posible que las operaciones admisibles no formen una lista tan bien definida como los medios permitidos en las demostraciones matemáticas —y aun aquí, ¿no podría alguien formular un medio *nuevo* aceptable para la demostración matemática que no cuadrara con ninguna regla previa conocida?—. A veces el problema consiste en divisar, crear o construir medios para alcanzar un objetivo, más que en emplear los medios existentes. (Puesto que la persona tendrá que usar otros medios para poder divisar éstos, ¿podría decirse que éste es el problema, mejor definido, al que se enfrenta?) Algunos medios de transformación pueden ser más costosos que otros, o ser sospechosos de yerro si se emplean con demasiada frecuencia. Una persona puede llegar a dudar incluso de si una operación es admisible o no. Al argumentar en contra de la posición mantenida por una persona, ¿está permitido apuntar al hecho de que esa persona no mantiene esa posición en su propia vida? ¿Se pueden usar en matemáticas argumentos de *reductio ad absurdum*? ¿Se pueden contar mentiras blancas (o grises o negras) para conseguir un objetivo valioso? Tampoco las restricciones están necesariamente bien definidas, y sus límites pueden estar poco claros. El pensamiento puede divisar un nuevo modo de evitar completamente una restricción

plemente con problemas dados; su tarea consiste en *fabricar* un problema, en *encontrar* un problema en las incipientes situaciones en que se hallan. Es posible que la tarea de formular un problema bien definido con objeto de captar los que a ustedes les parecerán los aspectos importantes de la situación en que se hallan no sea una tarea, el mejor modo de abordar la cual sea tratarla (o enfocarla) como un problema bien definido. Ello no obstante, el listado del modelo de problema con cuatro componentes (objetivos, estado inicial, operaciones admisibles y restricciones —el resultado es el producto que solventará el problema, no un componente constitutivo del mismo—) resulta iluminador, aunque quisiéramos relajar la especificidad del mismo para poder aplicarlo a situaciones de la vida real, y aunque pensemos que una parte importante de la actividad intelectual consiste no en resolver problemas, sino en llegar a ellos. No es necesario que un modelo ilumine todas las etapas de la actividad intelectual para que pueda iluminar una.

El modelo de problema describe cómo aparece la situación a la persona que se enfrenta al problema, y podemos usarlo para entender sus intentos de solución. Sin embargo, la persona puede estar equivocada respecto de determinados rasgos de su «espacio del problema»; puede tener más recursos de los que cree, menos restricciones de las que piensa. Aun así, su modo de entender la situación es lo que moldea lo que hace. En retrospectiva, podemos tratar de explicar por qué se equivocó acerca de algún particular aspecto del espacio del problema; y si la persona persevera, ella misma puede examinar si se ha equivocado en algún aspecto.

¿Cómo se *descubre* un problema intelectual, cómo se tiene noticia del mismo, cómo se aisla y se formula? Y una vez formulado el problema, ¿cómo se las arregla el pensador para darle una *solución* (o un intento de solución, al menos)? No tenemos ninguna teoría precisa de los *métodos* del trabajo intelectual, de los principios, las reglas tentativas, las máximas y las formas de construir esas obras intelectuales. La literatura pertinente llama *heurísticos* a los instrumentos que pueden ayudar a conseguir un resultado o solución, y los distingue de los algoritmos, que, en cambio, garantizan la pro-

---

dada, socavando o eludiendo las razones que parecían apoyarla. Las restricciones pueden ser una cuestión de grado, más o menos difíciles o costosas de violar. Haríamos mejor construyendo las restricciones como un *gradiente.* (¿Cuán *difícil* es moverse en esta dirección, cuánta energía hay que gastar para salir de esta situación o permanecer en ella, cuán costoso es, en términos de tiempo o de recursos, cuánta resistencia hay que vencer?)

ducción de la solución (si existe) en un número finito de pasos. La literatura formal de la heurística trata de formular reglas que pueden programarse en un ordenador y aplicarse mecánicamente. Nuestro propósito es diferente: hallar reglas y principios no mecánicos que sean de ayuda para una persona inteligente, aun si resultan necesarias cierta comprensión, inteligencia y pericia para aplicarlos.

Mi interés aquí es examinar reglas tentativas para la construcción de teorías filosóficas, o teorías intelectuales de otro tipo, para la formulación y la solución fértiles de problemas intelectuales. Pero el marco discutido aquí, de estructuras de problemas y de reglas heurísticas, puede resultar también útil para la inteligencia del modo en que (alguna) obra intelectual del pasado fue realizada. Discutir la «estructura del problema» como un foco organizativo capital para la historia intelectual nos distraería de nuestro interés principal aquí; añado una nota larga para el lector interesado.[43]

Karl Popper entiende la actividad de solventar problemas como un procedimiento que selecciona soluciones tentativas y las critica, modifica el problema a la luz de esa crítica, propone nuevas soluciones tentativas, y así sucesivamente, hasta que el problema queda resuelto. Lo que normalmente vemos, salvo en aquellos casos en que disponemos de cuadernos detallados de notas en los que el pensador plasma y desarrolla sus pensamientos,[44] es el *producto acabado*, que consiste habitualmente en la formulación de un problema (que puede no ser el problema original), una solución para este problema y quizás algunas réplicas a posibles críticas a esta solución. No es una tarea fácil la de reconstruir a partir del producto final el proceso entero, que acaso consista en muchos pasos repetidos que llevan a él. Los principios heurísticos entran en todas y cada una de las etapas. Podemos tratar de identificar principios o reglas para formular y seleccionar un problema, principios y reglas para formular una solución tentativa a un problema, principios para criticar posibles soluciones a un problema y principios para *re*formular un problema y modificarlo a la vista de las críticas y dificultades atraídas por las soluciones antes propuestas. El esbozo que de estas etapas hace Popper nos pone sobre aviso respecto de diferentes tipos de procedimientos heurísticos que podrían entrar en la construcción de un producto intelectual.

Un asunto que provoca cierta discusión actualmente es el de la medida en que los productos intelectuales resultan del uso de una heurística *general* que podría aplicarse a una amplia gama de ámbitos y objetos y el de la medida en que resultan de heurísticas específicas que incorporan abundante información acerca de estruc-

turas, pautas y procedimientos pertenecientes a un determinado dominio o subdominio intelectual.[45] También podríamos clasificar a los objetos de la investigación según el tipo de heurística, general o específica, que usen, lo mismo que podríamos, en parte, caracterizar el *estilo* de los distintos pensadores por la heurística que emplean.

He aquí algunos ejemplos de principios y procedimientos heurísticos —una muestra al azar—, casi todos relacionados con la primera etapa de esbozo de una solución tentativa para un problema intelectual. (Tengo presentes aquí problemas *teóricos;* para otros tipos de problemas intelectuales habría que aplicar otros tipos de heurística.) Esas *particulares* reglas tentativas son interesantes por sí mismas. También deberían servirnos para darnos cuenta de que el dominio de la racionalidad es mucho más amplio que el de los meros principios para evaluar la evidencia a favor y la evidencia en contra. Recuérdese que se trata de reglas heurísticas —no hay garantía de que vayan a tener éxito en ninguna aplicación particular—.

1. Cuando se ha dado durante mucho tiempo un conflicto entre posiciones intelectuales sin que se haya producido ningún movimiento importante ni se vislumbre solución, busca un supuesto o un presupuesto que sea común a *todas* las posiciones contendientes.[46] Trata de negar el supuesto y, en el espacio nuevo que abre esa negación, trata de construir una nueva posición.

Una posible explicación de la falta de soluciones intelectuales es que todas las posiciones y todas las partes contendientes han estado prisioneras de un marco intelectual que impide la adecuada solución del problema. (Otra explicación es que el marco actual es adecuado, pero nadie ha sido lo bastante listo como para resolver la cuestión en su seno.) Después de formular la solución, algún supuesto que todo el mundo daba por sentado parecerá algo arbitrario.

Mas ¿cómo puede identificarse un supuesto que todos, incluido uno mismo, comparten y dan por sentado? Considérese el problema de conectar los nueve puntos con cuatro líneas rectas sin levantar el lápiz del papel.* ¿Cómo puede identificarse el supuesto que excluye a la solución? Quizás haciéndolo todo aburridamente explícito —«cada línea trazada debe terminar en un punto»— y luego controlar y ver si el enunciado explícito está en realidad entre las condiciones originales del problema. (¿O sirve simplemente decir, para otros problemas, «recuerde que se le permite salir de los puntos»?)

* • • •
• • •
• • •

He aquí un supuesto que, según creo, da actualmente contexto a un problema. Cuando se dice, en relación con la paradoja de Einstein-Podoslky-Rosen y con la desigualdad de Bell, que la correlación entre las partículas separadas viola la localidad, se supone que la topología y la métrica del espacio se ha mantenido fija. Si la producción de dos partículas altera la topología y la métrica del espacio, generando, por ejemplo, a medida que éstas se separan, un creciente agujero en espiral entre ellas, entonces los efectos de una sobre la otra pueden ser (en este espacio remetrizado) bastante locales. Eliminar este supuesto de topología y métrica fijas, deja expedito el camino para investigar teorías alternativas, dentro de la geometría diferencial, que ofrezcan una estructura topológica adecuadamente distinta y para buscar y comprobar sus consecuencias contrastables.

2. Bastante a menudo, sospecho, un supuesto subyacente es explícitamente identificado sólo después de considerar una posibilidad radicalmente nueva. (De manera que la vía habitual de llegada a esta nueva posibilidad no pasa por empezar identificando el supuesto subyacente, para luego negarlo o suponer que no se da y, finalmente, ver qué nuevas posibilidades aparecen entonces.) Pero aún podemos usar esta posibilidad radicalmente nueva, una vez concebida, para preguntar cuál es el supuesto más profundo que viola, qué nuevo contexto resulta el adecuado una vez eliminado ese supuesto, y así sucesivamente.

3. Presta especial atención a la explicación de las simetrías o asimetrías inesperadas, a las simetrías o asimetrías que no tienen especial razón de existir o de no existir. (Einstein comienza su artículo sobre la relatividad especial diciendo de la electrodinámica de Maxwell, entendida del modo usual, que lleva a «asimetrías que no parecen inherentes a los fenómenos»).[47] Si se da una asimetría en una propiedad, pero todos los factores relevantes en un contexto se relacionan simétricamente con esta propiedad, considera la posibilidad de un contexto más amplio con objeto de descubrir un factor relevante que se comporte asimétricamente.

4. Aplica una operación o un proceso que ha sido fértil en otra parte a un caso nuevo que sea similar en los aspectos apropiados, modificando apropiadamente las diferencias entre los casos. (Por ejemplo, adopta el punto de vista de la producción aplicado por Emil Post a la generación del conjunto de haces de teoremas en lógica y aplícalo a la linguística, de manera que su objetivo sea el de generar el conjunto de haces gramaticales de una lengua —una vía por la que Noam Chomsky hubiera podido llegar a su inicial reestruc-

turación de la linguística.) Éste no es sino un caso de una máxima más general.

5. Ensaya modelos o analogías procedentes de otras áreas bien desarrolladas para estructurar el incipiente material con el que estás trabajando.

¿Cómo puede descubrirse una analogía *fértil* que nos ayude a resolver un problema en el área que es *blanco* del intento de estructuración? Suponiendo que ustedes ya han formulado un problema que tiene la estructura estándar ya descrita —objetivo, estado y recursos iniciales, operaciones admisibles y restricciones—, he aquí tres sugerencias procedentes de un trabajo sobre inducción.[48]

a) Empieza con el estado inicial del blanco y modifícalo sistemáticamente hasta llegar a una estructura intelectual que ya controlas. Usa las operaciones admisibles de *esa* estructura para formular operaciones análogas adecuadas al sistema-blanco, y aplícalas al estado inicial para ver si te llevan al objetivo. (Si te acercan a él, considera el modo en que estas operaciones análogas podrían ajustarse o modificarse para llevar exactamente al objetivo.)

b) Empieza con los objetivos *G* del blanco. Descubre objetivos estructuralmente similares *G'* en algún otro ámbito. Busca operaciones *O'* que lleven a *G'* en este otro ámbito. Formula operaciones análogas *O* en el ámbito del blanco. Controla si esas operaciones generan los objetivos *G* en el ámbito del blanco.

c) Empieza con los objetivos *G* del blanco y con el estado inicial del blanco. Formula una conjetura sobre cuáles serán los rasgos *F*1, ..., *F*n del estado inicial relevantes para alcanzar los objetivos. Da entonces una descripción *D* del estado inicial del sistema-blanco usando *sólo* esos rasgos. Busca otro ámbito con rasgos estructuralmente similares *F*1', ..., *Fn'*. Traduce los objetivos *G* del blanco a ese ámbito como *G'*. En ese ámbito, observa qué operaciones *O'* llevan del estado inicial (con rasgos *F*1', ..., *Fn'*) a *G'*. Traduce estas operaciones *O'* a la correspondiente operación *O* en el ámbito-blanco. Controla si *O* en el ámbito-blanco produce los objetivos *G*.

Éstos son *procedimientos específicos* para construir una analogía posiblemente fértil, procedimientos que podríamos usar para reconstruir los procesos intelectuales por los que transitaría (o podría haber transitado) un pensador. Otros procedimientos son menos específicos y resultarían problemáticos para un historiador que tratara de reconstruir la genealogía real de un producto intelectual. Por ejemplo: sumérjanse ustedes mismos en un problema y luego vaguen, busquen y estén alertas respecto de pistas que sugieran analogías fértiles. (Siguiendo este consejo, una persona podría ramonear por

entre los títulos de libros en su estantería, hojearlos, pasear por las calles, o cualquier otra cosa.) Lo que se presume aquí es que, cuando ustedes andan inmersos en un problema, las ayudas externas pueden producir chispas que apunten a ámbitos fértiles en los que buscar analogías. Obsérvese que, así como trabajar en un problema ayuda a preparar la mente para advertir soluciones, así también podría ser de ayuda atravesar los tres procedimientos explícitos antes mencionados de búsqueda de analogías. Y aunque no funcionen, le darán a uno una idea mejor del *tipo* de analogía que se necesita, y así, le ayudarán a orientar (inconscientemente, en parte) la búsqueda.

Tanto más capaz se será de descubrir analogías fértiles para un problema, cuanto mayor sea el fondo de diferentes estructuras intelectuales y teorías de que se dispone. (Para un problema difícil que se ha resistido a los intentos de otros, un instrumental de estructuras y problemas solventados de este tipo debería incluir ejemplos procedentes de las áreas más dispares. Otros ya habrán intentado sin éxito usar el equipo procedente de ámbitos más cercanos. Así, los formuladores de nuevos enfoques a menudo tienen una biografía rica en adquisiciones de recursos inusualmente variados.)[49] La mayoría de pensadores tienden a vivir del capital tempranamente adquirido; en algún momento, parece deseable añadir estructuras y herramientas difíciles nuevas que, sobre sus usos directos, pueden servir como fuentes de analogías y estimular la imaginación estructural.

6. Trabaja hacia atrás desde el objetivo y hacia adelante desde el estado inicial para ver si consigues que esta vía férrea transcontinental se encuentre.[50]

7. Reduce un problema difícil a un conjunto de problemas más fáciles, y usa otra heurística para resolver estos últimos.[51]

8. Examina casos *extremos*, considera qué pasa si algunos parámetros se hacen valer cero o infinito, y luego reconsidera tu caso intermedio a la luz de esta conducta extrema.

9. Investiga y enumera los rasgos generales que debe tener una solución correcta del problema. Busca algo que tenga esos rasgos. Si encuentras algo así, uno así, pero no más, trata de probar que este objeto es el único que satisface las condiciones y que, por lo tanto, es la solución. Si no puedes hallar ninguno, trata de probar que nada puede satisfacer todas esas condiciones; si tienes éxito en esta demostración, habrás obtenido un resultado de imposibilidad como los que a veces encontramos en la teoría de la elección social o en la teoría de la decisión en condiciones de incertidumbre.[52] La mera eliminación de una condición puede restaurar la consistencia,

pero dejar a las restantes condiciones en situación de ser satisfechas demasiado fácilmente por demasiados objetos diferentes. Relajar ligeramente una condición podría restaurar la consistencia preservando un conjunto firme de condiciones con una única solución —un resultado deseado—. La tarea consiste en estudiar cuál de las condiciones propuestas como solución debería modificarse, o cuál eliminarse y con qué substituirse luego, e investigar a continuación si algo satisface el nuevo conjunto de condiciones reformuladas.

10. Con una nueva idea particular, construye una pequeña estructura o un modelo formal para dar cauce a esa idea, y luego explora sus propiedades e implicaciones.[53]

11. Halla una descripción más *abstracta* de un proceso, noción o fenómeno e investiga sus propiedades para obtener un resultado más general y más potente; sigue aumentando la abstracción de la descripción hasta que los resultados sean menos potentes.

12. Al investigar una relación *R* (por ejemplo, «explica» o «justifica»), estudia también la estructura del conjunto del dominio inducido por *R*. ¿Qué problemas especiales plantea esta estructura global, y qué modificaciones de *R* producirían una estructura global diferente y mejor?

13. Transforma los fenómenos conocidos para descubrir fenómenos nuevos.[54] Cuando se sabe que *a* está en la relación *R* con *b*: (1) Investiga el *alcance* del fenómeno: ¿cuál es la lista de cosas que pueden substituir a *a* y sin embargo mantenerse en la relación *R* con alguna otra; cuál es la lista de cosas que pueden substituir a *b* y sin embargo mantenerse en esa relación *R*? (2) *Caracteriza* la dimensión del fenómeno describiendo las propiedades que acotan cada alcance. (3) Investiga cómo cambian las cosas si substituyes a *R* por otra relación aparentemente similar, *R'*. (4) Investiga qué fenómenos nuevos resultan de substituir en un proceso particular uno de los componentes por otro.

14. Si tratas de forzar una decisión o una descripción en un caso confuso exponiendo un caso análogo en el que la decisión o la descripción resulta clara, formula la diferencia que hace distintos a los dos casos y muestra por qué esta diferencia no basta para decidir los casos de modo diferente, ni menos opuestamente.[55]

15. Recientemente, ha aparecido un nuevo procedimiento para generar cuestiones, desentrañar embrollos y estimular ideas detalladas: construye una simulación por ordenador del fenómeno o proceso.

16. También sería útil formular algunos principios que sirvieran para provocar *experimentos intelectuales* en ciencia y en filosofía.

(Recuérdense el antropólogo de W.V. Quine, que hacía traducciones radicales, la tierra gemela de Hilary Putnam, los constructores de Ludwig Wittgenstein y el ejemplo de la máquina de experiencias.)[56]

Creo que vale la pena ensayar estos particulares principios heurísticos. Al consignarlos aquí espero también estimular a otros —no sólo a los filósofos— a formular principios heurísticos fértiles para las varias etapas de la investigación intelectual. No principios «mecánicos» —su uso puede presuponer muchos conocimientos y «sensibilidad» respecto al material—. Resulta un poco sorprendente que, al entrenar a los estudiantes de filosofía, los profesores no hagan ningún esfuerzo serio para formular reglas tentativas de este tipo para conducir la empresa intelectual. Los estudiantes sólo reciben productos acabados (libros y artículos) y el ejemplo de sus profesores como analistas y comentadores de los mismos, para ser luego abandonados a representarse por sí mismos el modo en que estos productos pudieron ser realizados. No vendrían mal algunas pistas explícitas.

La imaginación de la racionalidad

Los procedimientos y pautas instrumentales del pensamiento teórico racional son variados. Al pretender cosas distintas de la creencia —por ejemplo, productos intelectuales nuevos y fértiles—, no siempre se centran en disciplinar y evaluar las razones a favor y en contra (los principios heurísticos constituyen un ejemplo de ello). Tampoco la racionalidad de la creencia se reduce a una cuestión de aplicar reglas (mecánicas de ponderación de razones dadas. La imaginación desempeña un papel importante. Entiendo por imaginación simplemente la capacidad para pensar en posibilidades nuevas y fértiles. El listado de reglas heurísticas, pues, podría no ser sino el comienzo de una teoría que definiera qué es esa imaginación y cómo opera.

La credibilidad de un enunciado viene determinada por las razones a favor y en contra y por otros enunciados que socavan o dan fuerza a esas razones. Sin embargo, detectar posibilidades alternativas que podrían socavar a un enunciado no es un asunto mecánico. Se presenta evidencia en favor de una hipótesis, pero ¿se han controlado todas las variables relevantes? Cada variable relevante apunta a una hipótesis alternativa que podría dar cuenta de los datos. Por eso deben controlarse todas (de otro modo, el peso del apoyo conferido a la hipótesis por los datos existentes será menor). Pero

la imaginación y el ingenio resultan necesarios para detectar qué variables olvidadas podrían desempeñar plausiblemente un papel en la aparición de los datos. En la aplicación de la regla 1 para la creencia también entra la imaginación. ¿Algún enunciado competitivo e incompatible tiene un valor de credibilidad mayor? Formular el enunciado alternativo más digno de consideración no es un asunto mecánico —la teoría de la relatividad es una alternativa a la mecánica newtoniana, pero sólo Einstein consiguió formularla—; tampoco lo es a veces saber que se trata de una alternativa, de un enunciado en conflicto e incompatible.

La producción de alternativas nuevas juega un papel importante tanto en la acción como en la creencia. La elección de una acción se hace entre alternativas. Elegir mejor entre las alternativas existentes es una manera de mejorar los resultados. Otra manera de hacerlo es ampliar la gama de alternativas para incluir en ella nuevas alternativas prometedoras. La construcción imaginativa de una nueva alternativa, no concebida hasta ahora, podría ser condición de posibilidad de la mayor mejora posible. Podría haber reglas útiles sobre la oportunidad de buscar alternativas nuevas de este tipo, pero los resultados dependerán de que las encontremos realmente. No hay ningún procedimiento mecánico (algorítmico) para generar las alternativas más prometedoras —no que sepamos, al menos—. Las máximas heurísticas podrían venir en nuestra ayuda.

¿Convierten estas consideraciones a la imaginación en un componente de la racionalidad? Algunos podrían insistir en que la racionalidad consiste sólo en realizar la mejor elección entre las alternativas *dadas* —acciones o creencias—. Me parece una ablación innecesaria y arbitraria. Mas, aun si la imaginación no fuera un componente, sí sería al menos una compañera de la racionalidad, una compañera importante como medio para conseguir los objetivos de la racionalidad. En algunas situaciones, podría ganarse mucho más generando nuevas alternativas y eligiendo aproximadamente entre ellas que limitándonos a elegir finamente y con perfecta discriminación entre las alternativas existentes. El óptimo de segundo grado entre las alternativas nuevas podría ser muy superior al óptimo entre las alternativas viejas. Es tan importante cultivar las facultades imaginativas relevantes como aguzar las facultades discriminantes.

Sin la exploración y la comprobación de otras posibilidades imaginativas, los procedimientos de la racionalidad, centrados exclusivamente en las alternativas *dadas*, serán miopes. Aun haciéndonos bien, pueden reducirnos a un óptimo local. Una conocida analogía

empleada en la literatura sobre maximización relaciona a ésta con la actividad de alcanzar el punto más alto en un terreno geográfico. Una persona corta de vista, con un alcance de visión de diez pies, podría seguir este procedimiento: escrutar los alrededores exhaustivamente, y luego moverse hasta el punto más alto que pueda ver (que estará dentro de los diez pies); repetir de nuevo el procedimiento una y otra vez hasta que se llegue a una posición en la que ningún punto visible sea más alto que el sitio en el que se está; detenerse entonces. Si esta persona sale de la ladera de una loma, este procedimiento le llevará a la cumbre; pero no le llevará a la cumbre de la loma más alta siguiente, de cuya ladera no partió. Este procedimiento le llevará al punto local más alto, a un óptimo local, pero no al punto absolutamente más alto, al óptimo global.

Sin la producción y la comprobación imaginativas de las nuevas posibilidades, la racionalidad sola nos llevará únicamente a un óptimo local, a la mejor de las alternativas ya dadas. Es un rasgo estimable de la racionalidad el que sea capaz de *conducirnos* hasta allí. Pero debemos guardarnos de que la racionalidad no opere de tal forma que nos *restrinja* a eso. Es demasiado fácil y tentador para la racionalidad el convertirse en un mecanismo que clasifique como irracionales y excluya las actividades de generar y comprobar imaginativamente posibilidades nuevas. El proceso de exploración de posibilidades nuevas será imperfecto y aparentemente despilfarrador; muchas posibilidades exploradas acabarán siendo inútiles. Pero la racionalidad debe ser tolerante al respecto, y no exigir de antemano garantías de éxito.

Los contextos de descubrimiento y de justificación —concebir hipótesis y estimar su credibilidad— no pueden separarse completamente. Para estimar la credibilidad de una hipótesis, tenemos que concebir y estudiar su mejor alternativa incompatible. (Podríamos requerir explícitamente que la estimación de una hipótesis se hiciera en relación con una hipótesis alternativa dada, pero eso no nos proporcionaría ninguna conclusión deslindable respecto de la primera hipótesis.)

La cuestión «¿qué nuevas posibilidades alternativas hay?» es el primer paso en el progreso humano, en la producción de nuevas teorías, de nuevos inventos, de nuevos modos de hacer, de actuar, de cooperar, de pensar y de vivir. Plantearse esta cuestión presupone una disposición a romper con la tradición, a aventurarse en territorio desconocido. Responderla presupone la capacidad para concebir posibilidades nuevas y fértiles; es decir, presupone imaginación.

No todo el mundo querrá explorar posibilidades en todos los ám-

bitos, y sería ineficiente que todo el mundo lo intentara. Nos beneficiamos de las actividades de los demás y de las diferencias entre nosotros que llevan a otros a hacer, y a pensar, lo que nosotros no haríamos y no pensaríamos. Los filósofos de la ciencia han tratado de formular un procedimiento mecánico que llevara a todos los científicos a tomar exactamente la misma decisión acerca de si hay que aceptar una teoría científica determinada (en una situación en la que se dispone de determinada evidencia se han formulado determinadas teorías alternativas, etc.). Pero las diferencias de opinión cumplen una importante función en el progreso corriente de la ciencia, como observó Thomas Kuhn. Las cosas van bien cuando algunos científicos exploran y propugnan vigorosamente teorías nuevas, mientras que otros defienden y reforman con igual vigor la teoría existente para que dé cuenta de fenómenos nuevos. El desarrollo a lo largo de estas diversas avenidas es lo que acaba produciendo el conocimiento detallado de las capacidades y limitaciones de las distintas teorías, y así, generando el acuerdo general que los científicos acostumbran a mostrar.[57]

Frederick Hayek resaltó cómo también la vida social se beneficia de las exploraciones llevadas a cabo por individuos que ensayan nuevos métodos de producción, que desarrollan nuevos productos, que experimentan nuevas pautas de conducta y modos diferentes de vida, de manera que todos salimos ganando de que haya libertad general para explorar de esta forma, aun cuando nosotros mismos no la aprovechemos. Producidos los primeros resultados de las exploraciones de algunos, otros pueden elegir emularles, y finalmente nosotros podemos seguir el camino también. Aun si no lo hacemos, podemos beneficiarnos de las actividades de quienes lo hacen, y por lo tanto, de las actividades de los primeros exploradores e innovadores. Quizá la mayoría de esas exploraciones e innovaciones sean infértiles, pero los costes los arrostrarán quienes hayan elegido explorar; y cuando se difundan los efectos de las (relativamente) escasas innovaciones fértiles, todos saldremos beneficiados.[58] La naturaleza social de nuestra vida económica, intelectual y política nos permite beneficiarnos de imaginaciones que nosotros mismos no tenemos —y nadie puede ser igualmente imaginativo en todos los ámbitos, aunque sólo sea porque esto requiere una atención y un grado de alerta que sólo limitadamente poseemos—.[59]

La racionalidad tiene dos caras. Lo que resulta estimulante de la racionalidad es su aguzado filo, su temeraria intrepidez. Los exigentes criterios de credibilidad y la plena consideración de las contrarrazones y de los factores socavadores hacen de la regla 1 un arma

potente; no creas algo cuando resulta más creíble algún enunciado incompatible. Los lectores jóvenes se estremecen, como se estremecieron sus jóvenes oyentes, con la exactitud y el coraje de los aguijonazos infligidos por Sócrates a la pompa de la creencia vanidosa. La intrepidez de la racionalidad consiste en su disposición a formular lo que antes no estaba al alcance del pensamiento y en defender una creencia previamente descartada como ultrajante —siempre que, consideradas todas las razones, *sea* en efecto más creíble que cualquier competidora—. Tampoco en lo que hace a las decisiones respeta la racionalidad restricción arbitraria alguna; la cuestión es *qué* acción maximiza todas las funciones relevantes, y el que esa acción sea inaudita no cuenta para nada. Esta gloria de la racionalidad se revela del modo más claro en las enigmáticas, llamativas, sorprendentes y a veces desconcertantes teorías producidas por la investigación científica, pero también se revela en contextos más cotidianos. La primera cara de la racionalidad es la cara romántica, dibujada por la estimulante narración popperiana del decurso de la ciencia como la crítica y la comprobación despiadadas de teorías cada vez más intrépidas, siempre a lomos de la precaria cuerda funambulesca tendida sobre el abismo de la refutación. Aun conociendo sus insuficiencias y sus lagunas, esa narración resulta inspiradora.

Pero la potencia de la racionalidad no tiene exclusivamente que ver con sus llamativos triunfos aisladamente considerados. La racionalidad tiene una fuerza acumulativa. Una decisión racional dada quizá no sea mucho mejor que una decisión menos racional, pero lleva a una nueva situación de decisión algo distinta de la situación que habría ocasionado la otra decisión y en esa nueva situación de decisión, una ulterior decisión racional produce sus resultados y lleva a otra situación de decisión. Con el tiempo, pequeñas diferencias de racionalidad se amalgaman para producir resultados muy diferentes. Una creencia dada acaso no sea mucho más creíble que una creencia incompatible con ella, pero esas dos creencias vienen en apoyo de dos enunciados diferentes, y la amalgama de diferencias de credibilidad que se va produciendo a lo largo de estas vías puede llevar a cuerpos de creencia inmensamente diferentes. En el ajedrez, algunos jugadores acaban doblegando al adversario a través de una acumulación de pequeñas ventajas; otros jugadores realizan ataques y sacrificios tan osados como pertinentes. Como los más grandes campeones de ajedrez, la racionalidad hace ambas cosas.

Nuestras exploraciones nos han llevado a nuevos principios de racionalidad. Un principio de decisión racional exige la maximiza-

ción del valor decisional, lo que nos lleva más allá de la estructura meramente instrumental de la racionalidad. Dos principios gobiernan la creencia racional (incluso en la apariencia puramente teórica), disolviendo el dualismo entre racionalidad teórica y práctica; no creas ningún enunciado menos creíble que alguna alternativa incompatible —el componente intelectual—, pero cree un enunciado sólo si la utilidad esperada de creerlo es mayor que la de no creerlo —el componente práctico—. Y la racionalidad de la creencia entraña dos aspectos; apoyo por parte de razones que hagan creíble la creencia, y generación por parte de un proceso que fiablemente produce creencias verdaderas. Nuestro enfoque evolucionario de las razones explica la enigmática conexión entre esos aspectos, pero invierte la dirección de la «revolución copernicana» de Kant.

Esta perspectiva evolucionaria arroja también una imagen nueva de la naturaleza y del estatus de la racionalidad, la cual es una característica destinada a jugar un importantísimo papel en la determinación de aquello que hay de especial en la condición humana. La racionalidad es una adaptación evolucionaria con un propósito y una función delimitados. Fue seleccionada positivamente y diseñada para trabajar en equipo con hechos duraderos que se mantuvieron constantes a lo largo de la evolución humana, independientemente de que los hombres pudieran o no demostrar esos hechos. Muchos de los problemas filosóficos tradicionalmente indómitos, resistentes a su resolución racional, acaso resulten de intentos de extender la racionalidad más allá de esa función delimitada. Entre esos problemas está el de la inducción, el de las otras mentes, el del mundo externo y el de la justificación de los fines —el intento kantiano de convertir la conducta regida por principios en el único y último criterio de conducta no es sino otra extensión de la racionalidad más allá de sus límites—. Sería un accidente increíblemente afortunado —pero no imposible— el que una racionalidad como la nuestra, modelada para otros propósitos delimitados, bastara a demostrar la verdad de *todas* las condiciones con las que coevolucionó para trabajar en equipo con ellas.

Hemos explorado los límites de la racionalidad instrumental, pero no deberíamos poner demasiado énfasis en ellos. La racionalidad instrumental es una herramienta poderosamente disciplinada que permeará, y a la que tendrá que incluir, como parte significativa de ella misma, cualquier otra concepción de la racionalidad que aspire a ser completa. Las condiciones que hemos puesto a la racionalidad de los objetivos sugieren una noción ampliada de racionalidad instrumental en tanto que búsqueda efectiva y eficiente de

objetivos racionales. Esas condiciones, empero, no determinan plenamente la racionalidad substantiva de los objetivos y los deseos, y quizá debamos estar agradecidos, por eso. Una teoría plenamente determinada de la racionalidad substantiva abre las puertas para exigencias despóticas, externamente impuestas. Es verdad que la falta de esa teoría permite que haya algunos deseos objetables, incluidos algunos deseos faltos de ética, pero quien debería lidiar con éstos es una teoría ética substantiva —a pesar de los persistentes intentos de los filósofos por subsumir la ética en la racionalidad—. La racionalidad instrumental nos deja el espacio necesario para perseguir autónomamente nuestros *propios* fines.

Por lo demás, hemos desarrollado una noción de racionalidad que incluso va más allá de la noción ampliada de lo instrumental (como la persecución efectiva de objetivos racionales), para incluir lo simbólico y lo evidencial. Es verdad que nuestra consideración de los factores simbólicos y evidenciales, lo mismo que la de los factores causalmente instrumentales, puede tener también un origen evolucionario. Cualquiera que sea la función evolucionaria de nuestra capacidad simbolizadora —ya sea la de robustecer otros deseos, o la de mantenerles firmes a través de períodos carentes del refuerzo que les proporcionan sus objetos reales, o la de permitir a la gente coordinar sus acciones en situaciones de dilema del prisionero en las que no podría darse de otro modo la cooperación—, no es necesario que esa función se convierta en nuestro objetivo presente, como no se ha convertido en nuestro objetivo la maximización de la adaptación inclusiva. Una vez poseemos la capacidad, y cualquiera que sea su razón de ser, podemos utilizarla y ponerla al servicio de nuestras propias razones y de nuestros propios propósitos. Lo mismo que en el caso de nuestras capacidades matemáticas, no estamos obligados a reducir esa capacidad al cumplimiento de su función original.

La base evolucionaria de nuestra racionalidad no nos condena a proseguir ninguna pista evolucionaria previamente dibujada. (Pero tampoco el conocimiento de que la función evolucionaria de un rasgo ha dejado de servir garantiza que elijamos cambiar ese rasgo, aun siendo capaces de ello. Podemos conservar el rasgo porque fue el impulso hacia un modo de vida que ahora valoramos independientemente, y así, le damos una nueva función.) Nos hemos servido de nuestras capacidades racionales para obtener conocimiento de esta base evolucionaria, aunque nuestras capacidades no fueron precisamente seleccionadas para eso. Objetivos que fueron inculcados porque servían a la adaptación inclusiva pueden ahora ser perseguidos aun si están en conflicto con esa adaptación —perseguidos

por individuos o (durante algún tiempo al menos) por grupos—. Podemos usar nuestra imaginación para formular nuevas posibilidades, ya sean objetivos, teorías o planes osados, sin que ninguna de ellas arraigue en funciones evolucionarias específicas previas. Y aun si la imaginación misma, la capacidad para concebir nuevas posibilidades, tiene una función evolucionaria, ahora podemos usar esa capacidad para cualesquiera propósitos que elijamos.

Las creencias y las acciones de una persona racional están generadas (y mantenidas) por un proceso que consigue fiablemente determinados objetivos, un proceso en cuya orientación desempeñan cierto papel apropiado las razones. Lo que originariamente haya que contar como razón puede tener una base selectiva evolucionaria, pero la orientación proporcionada por las razones no es mecánica o ciega. Tomamos conocimiento de las razones, las ponderamos, consideramos objeciones a y posibles socavamientos de ellas, y moldeamos nuestra conducta y nuestras creencias de acuerdo con todo eso.

Ser autoconscientes de las razones y del razonamiento añade una nueva dimensión de control y desarrollo. La filosofía fue la primera disciplina que llevó esa autoconsciencia más allá de lo que es habitual entre las personas reflexivas en general, y lo hizo convirtiendo al razonamiento mismo en objeto de estudio. (Hegel y Fichte convirtieron luego a la autoconsciencia en objeto de estudio.) Se formularon propósitos y principios, se criticaron, se reformularon, se desarrollaron ulteriormente y se conectaron sistemáticamente entre sí. (Otros se han sumado luego a los filósofos en esta tarea, produciendo una abundante literatura de estadística teórica, teoría de la decisión y ciencia cognitiva.)

Este cuerpo desarrollado de principios teóricos —al que se ha llegado en parte merced al uso de principios que la evolución misma ha inculcado, pero rebasándolos— puede usarse para orientar nuestro pensamiento y nuestra conducta. Ese proceso llega también a ser autoconsciente. El grupo de principios de racionalidad que la gente desarrolla explícitamente puede aplicarse a esos mismos principios —algunos de ellos, a otros de ellos, y quizá algunos a sí mismos— generando modificaciones y desarrollos nuevos. Esa trayectoria puede alejarnos de nuestros orígenes evolucionarios.*

* Pero existe una posibilidad de que un defecto inicial de racionalidad pueda verse magnificado en vez de corregido cuando se usa a la racionalidad para modificarse a sí misma; ese defecto, al aplicarse, causa defectos mayores. Hay que andarse con cuidado para evitar esto, por ejemplo, empleando tests externos de tal tipo, que su veredicto no pueda ser automáticamente admitido como positivo, aun viniendo del sistema modificado de principios (presuntamente) racionales que está siendo sometido a prueba.

La racionalidad evolucionó como una adaptación sobre un trasfondo de hechos estables, con los cuales fue seleccionada positivamente para trabajar en equipo. Uno de esos hechos es la presencia de otras criaturas dotadas de una racionalidad evolucionada similar. Descartes nos pinta a un individuo reflexionando sólo en su estudio sobre cuál de sus creencias no podría ser falsa o producto de un engaño artero. Sus meditaciones nos ofrecen un procedimiento que ha de seguir también cada uno de sus lectores, solo en su propio estudio. Sin embargo, no hay razones para creer que la evolución ha modelado nuestra racionalidad de un modo acorde con este individualismo cartesiano. Si la racionalidad evolucionó junto con la concurrente racionalidad de los demás, entonces la racionalidad de cada persona tiene que ser de tal naturaleza que case con y se adapte al trabajo en equipo con la similar racionalidad de los demás.[60] No podríamos esperar entonces que la racionalidad pruebe que los demás son racionales, ni que sea capaz de probarlo; esto es algo de lo que parte la racionalidad y con lo que trabaja para hacer otros negocios.

¿De qué modos usa nuestra racionalidad la racionalidad de los demás? Estamos predispuestos a aprender de los demás el lenguaje y estamos predispuestos también a registrar hechos que nos muestran o nos cuentan nuestros mayores. Estamos predispuestos a aceptar lo que dicen y a aceptar sus correcciones de lo que dicen, al menos hasta que hemos atesorado suficiente lenguaje e información como para ser capaces de albergar dudas fundadas y de plantear cuestiones. «¿Pero por qué es racional confiar siempre en cualquier cosa que les diga otra persona?» Nuestra racionalidad no está construida para responder a esta cuestión; está más bien construida sobre esa confianza, y sobre ésta construye ella misma aún más. Si confianza es esto (es posible que esto se parezca más a aceptar sin reflexión lo que otros —o el primer grupo que conocemos— nos enseñan).

Una vez adquirida cierta base de lenguaje y de creencias fácticas, podemos usarla para cuestionar y modificar las creencias de otros. La común evolución de la racionalidad no condena a nadie al conformismo intelectual. Si hay una presunción de entrada en favor de creer a los demás, esa presunción puede ser superada. ¿Estaban éstos en una posición que les permitiera creer racionalmente un particular enunciado o recibirlo de alguien que estaba en esa posición? ¿Es esta persona miembro de un grupo, de una clase adecuada de referencia, tal que hay estadísticas sobre las creencias del grupo sobre un asunto y esas estadísticas minan nuestra presunción

en favor de la creencia? ¿Hay alguna razón especial para pensar que esta persona está motivada para confundirles a ustedes o para no esmerarse en ser exacta? ¿Han pensado ustedes en posibilidades en las que no ha pensado la otra persona y que son relevantes para evaluar la creencia? En este ámbito interpersonal también hay lugar para la formulación, discusión y desarrollo de principios adicionales para la creencia racional.[61]

El lenguaje es una manifestación y un vehículo de la racionalidad, y su naturaleza social ha sido puesta de relieve por muchos autores: Wittgenstein habla del papel del acuerdo en los juicios; Quine escribe sobre el lenguaje como arte social; y Hilary Putnam esboza la división del trabajo linguístico, merced a la cual la referencia de algunos de nuestros términos está determinada por el modo en que nos apoyamos en el conocimiento de los expertos.[62] Puesto que nuestras capacidades linguísticas evolucionaron en equipo con las de los demás, todos los cuales nacieron en un ambiente de hablantes adultos —dejemos ahora de lado las especulaciones sobre el origen del lenguaje—, resultaría sorprendente que los fenómenos del lenguaje y del significado fueran independientes de esos entornos sociales.

La evolución en un medio poblado por otras personas hace posible la especialización en rasgos diseñados para funcionar bien en presencia de otros rasgos. La innata «propensión a traficar, trocar y cambiar una cosa por otra» de que habló Adam Smith[63] serviría de poco si sólo la tuviera una persona. Una propensión al intercambio requiere compañeros con similares propensiones. Y así como hay una división del trabajo y una especialización de los talentos en la sociedad, ¿podría haber así también una división de las características biológicas dentro de un grupo, con alguna gente más belicosa y marcial, algunos más diligentes, algunos más sagaces y algunos más fuertes, dado que no es biológicamente factible que los humanos tengan todos esos rasgos y dado que muchos de ellos, si no todos, se beneficiarán viviendo con otros que tienen rasgos complementarios? (Por mi parte, para explicar esa variedad, preferiría hallar una base selectiva individual, no de grupo.)

Esto parece plausible, pero hay otra idea más turbadora. ¿Podría la racionalidad misma no ser sino un conjunto de rasgos que mostrara cierta variedad natural entre los miembros del grupo? En la sociedad de cazadores y recolectores del Pleistoceno ¿se dio una selección evolucionaria que hizo que algunos fueran más intensamente racionales en la creencia y en el cálculo, como la hubo para que unos fueran más fuertes y algunos más diligentes? ¿Se beneficia-

ron todos de las actividades cooperativas y de los intercambios que esa mezcla de rasgos propició? Pero acaso esa variación en la racionalidad revelada se debe únicamente a diferencias no biológicas. Cualquiera que sea la causa, se puede sospechar que los más racionales hace tiempo que se han mostrado acreedores a un estatus fortalecido, aunque sólo fuera porque eran mejores a la hora de articular y defender verbalmente sus exigencias. Otros, sin duda, prestan menos atención a este autoacicalamiento del intensamente racional.

Lo que ha acontecido, empero, es que la racionalidad ha remodelado al mundo. Éste es el gran tema de los escritos de Max Weber: el cálculo económico y monetario, la racionalización burocrática, las reglas y los procedimientos generales han venido a reemplazar a la acción fundada en vínculos personales, y las relaciones de mercado se han ido extendiendo a nuevos ámbitos.[64] La racionalidad, junto con los correspondientes cambios institucionales que utilizan explícitamente y dependen de la racionalidad, ha proporcionado muchos beneficios y ha permitido así a la racionalidad seguir extendiendo su dominio.

No obstante, esto ha hecho que el mundo sea ahora, en varios sentidos, poco hospitalario para grados menores de racionalidad. A aquellas culturas cuyas tradiciones no son receptivas a la racionalidad weberiana les ha ido menos bien. En las sociedades occidentales, el equilibrio en la división de rasgos que fue útil en las sociedades de cazadores y recolectores se ha desplazado. Al principio, la racionalidad fue capaz de extender su influjo porque también trajo beneficios para los demás rasgos, pero los demás rasgos se hicieron más dependientes de la racionalidad, y la racionalidad se hizo más poderosa y menos sujeta a restricciones. La racionalidad está ahora procediendo a rehacer el mundo a su conveniencia, modificando no sólo su propio medio, sino también el medio en el que todos los demás rasgos se hallan, extendiendo el medio en el que sólo ella puede florecer plenamente. En ese medio, el producto marginal de la racionalidad crece; el de los otros rasgos, disminuye; rasgos que antaño tenían una importancia coordinada, se sitúan ahora en una posición subordinada. Esto presenta un reto a la compasión de la racionalidad, y a su imaginación y a su ingenio: ¿puede ella concebir un sistema en el que quienes tengan otros rasgos puedan vivir confortablemente y florecer —teniendo, si lo eligen, la oportunidad de desarrollar su racionalidad—? ¿Y querrá?

Platón habló de aprehender las formas eternas, Aristóteles, de la intuición intelectual de los primeros principios y de las esencias conocedoras de la mente. Descartes, de las ideas claras y distintas

y de verdades bañadas en la luz natural de la razón, y Spinoza, de la cognición intuitiva de la esencia de las cosas. El lector de nuestro enfoque evolucionario bien podría preguntarse qué se ha hecho de la Dignidad de la Razón. Los enfoques deflacionarios de la razón no son nuevos; también Hume y Kant asignaron a la razón funciones más modestas. (Kant privó a «la razón especulativa... de sus pretensiones de visión transcendental», dijo, «para poder hacer espacio para la fe».) Sin embargo, el conocimiento de los orígenes y las funciones originales de la razón no le roba a ésta toda su nobleza. (Haremos bien en recordar que tampoco la nobleza, a pesar de sus frecuentes protestas, tiene origen especial alguno.) Considérese la curiosidad. Aun si algún grado de curiosidad fue seleccionado positivamente debido a su papel en la adquisición de nuevas verdades útiles prácticamente, una vez existe esa capacidad puede desviarse para investigar los orígenes del universo, la naturaleza del infinito, los orígenes y el desarrollo de la vida en la tierra y el alcance y los límites de la razón humana, todo lo cual con el único objetivo de satisfacer la curiosidad intelectual misma y por el conocimiento que trae consigo, sin ningún otro motivo o propósito ulterior. Y si la razón no fuera un infalible conocedor de una realidad independiente, quizá eso no haría sino resaltar lo impresionante y sorprendente de sus triunfos. Cualesquiera que hayan sido los orígenes prácticos del discernimiento estético, el caso es que ha sido usado para producir grandes obras de arte. Cuando nos damos cuenta de que las más sublimes creaciones humanas proceden de orígenes y funciones humildes, lo que tenemos que revisar no es nuestra estima por esas creaciones, sino nuestra noción de nobleza. ¿Hasta qué punto *fue* humilde un punto de partida que pudo impulsarnos hasta los logros humanos más sublimes, hasta qué punto fue mediocre algo que dio origen a tamañas potencialidades y potencias?

La racionalidad nos proporciona mayor conocimiento y mayor control sobre nuestras propias acciones y emociones y sobre el mundo. Aunque nuestra racionalidad es, inicialmente, una cualidad evolucionada —la naturaleza de la racionalidad incluye en ella a la Naturaleza—, nos permite transformarnos a nosotros mismos y, por lo tanto, transcender nuestros estatus como meros animales, realmente y también simbólicamente. La racionalidad viene a modelar y a controlar su propia función.

Nuestros principios fijan aquello por lo que vale nuestra vida, nuestros propósitos crean la luz en la que nuestra vida se baña, y nuestra racionalidad, así la individual como la coordinada, define y simboliza la distancia que hemos recorrido desde la mera anima-

lidad. Por esos medios nuestras vidas llegan a significar más que lo que instrumentalmente dan. Y al significar más, nuestras vidas dan más.

NOTAS

NOTAS AL CAPÍTULO 1

1. Un supuesto más débil mantendría que no *todo* juicio correcto surge de un principio aceptable, pero sí algunos o la mayoría de ellos. Con todo, encontrar un principio general aceptable que arroje un juicio particular tendería a mostrar que el juicio era correcto. Sin embargo, el fracaso a la hora de encontrar un principio así no constituiría una razón concluyente para abandonar el juicio, pues podría tratarse de un juicio del tipo de los juicios solitarios, que no son consecuencia de principio aceptable alguno.

2. Mark Tushnet ha argüido que en el ámbito jurídico el requisito de la discusión guiada por principios no constituye una restricción al resultado al que el juez puede llegar; si los casos previos casan con un principio (incluso con un principio firmemente establecido), cuyo resultado en el caso presente es considerado vitando por el juez, entonces este caso siempre podrá ser distinguido de los demás por algún que otro rasgo. Véase Tushnet, «Following the Rules Laid Down: A Critique of Interpretavism and Neutral Principles», *Harvard Law Review* 96 (1983): 781-827. Sin embargo, limitarse a distinguir el caso permite (cuando mucho) formular un juicio nuevo; no significa un *apoyo* para ese juicio. Para apoyarlo, el juez debería formular un nuevo principio cuyo tenor literal resultara plausible y que casara con (la mayoría de) los viejos casos, con el caso nuevo y con algunos casos hipotéticos obvios. Es decir, para proceder a la distinción que pretende, y para que esa distinción conduzca a una diferencia, el juez necesitará una justificación razonable en términos de principios. No es cosa fácil formular principios aceptables, y mucho menos hacerlo con la frecuencia que uno desearía cuando se le presentan casos nuevos que lo solicitan.

3. Véase C.G. Hempel, *Aspects of Scientific Explanation* (Nueva York: Free Press, 1965) [trad. cast.: *La explicación científica. Estudios sobre la filosofía de la ciencia*, Barcelona, Paidós, 1988], págs. 264-272, y Ernest Nagel, *The Structure of Science* (Nueva York: Harcourt, Brace and World, 1961) [trad. cast.: *La estructura de la ciencia*, Barcelona, Paidós, [3]1991], págs. 47-78.

4. El razonamiento abstracto guiado por principios da apoyo a una particular posición reclutando, en calidad de apoyo, otros juicios aceptados. Algunos autores han sugerido que este mecanismo impersonal es un mecanismo particular de justificación.

5. No me he detenido a averiguar qué posibles estudios empíricos de las decisiones de la gente pueden apoyar esta tesis *empírica* de los teóricos del derecho, ni qué estructura jurídica alternativa funcionaría como elemento de control experimental, etc.

6. El pequeño conjunto de datos de puntos que tenemos parece caer en una línea recta, pero los fenómenos emparentados que hemos descubierto apuntan a que no se da una relación lineal. Quizá se trata, pues, aquí de un accidente de los datos particulares que nos ha sido dado conseguir.

7. Véase la lista de factores en Thomas Kuhn, «Objectivity, Value Judgment and Theory Choice», en su *The Essential Tension* (Chicago: Univ. of Chicago Press, 1977) [trad. cast.: *La tensión esencial*, Madrid, FCE, 1983], págs. 320-339, y W.V. Quine y Joseph Ullian, *The Web of Belief*, 2.ª ed. (Nueva York: Random House, 1978), págs. 64-82. La necesidad de esos criterios adicionales no necesariamente viene de la finitud de nuestros datos. Quine ha sostenido que la totalidad de las observaciones posibles no basta para seleccionar unívocamente una teoría explicativa. (Véase su «On the Reasons for Indeterminacy of Translation», *Journal of Philosophy* 67 [1970]: 178-183, y «On Empirically Equivalent Systems of the World», *Erkenntnis* 9 [1975]: 313-328.) A falta de una teoría adecuada de la explicación y de lo que podría entrañar el detalle de la estructura de la relación explicativa, resulta difícil establecer la verdad de esta tesis tan fuerte.

8. Véase P. Atiyah y R. Summers, *Form and Substance in Anglo-American Law: A Comparative Study in Legal Reasoning, Legal Theory, and Legal Institutions* (Oxford: Oxford University Press, 1987).

9. Sería interesante investigar hasta dónde llega el paralelismo entre los rasgos estructurales de las razones y los de las causas y poder explicar por qué se da ese paralelismo. ¿Corren las razones en paralelo a fenómenos de causalidad probabilista?

10. Véase Herbert L.A. Hart, *The Concept of Law* (Oxford: Clarendon Press, 1961), págs. 155-159, y Chaim Perelman, *The Idea of Justice and the Problem of Argument* (Londres: Routledge and Kegan Paul, 1963).

11. El hecho de que otros puedan contar con que nosotros secundemos determinados principios también podría disuadirles de algunas acciones en vez de inducirles a cooperar. Una nación o una persona que albergue un principio de represalias contra determinadas ofensas, independientemente de los intereses inmediatos, podría disuadir a otros de realizar ofensas de ese tipo. Anunciar un principio así incrementa el coste de hacer excepciones para garantizar que no se hará nada.

12. El gobierno de los Estados Unidos desea emitir deuda y promete no incrementar la inflación de su moneda, pero, tras ser adquirida la deuda por otros, el gobierno tendrá interés en promover la inflación —y esos otros lo percibirán de antemano—. De aquí que el gobierno trate de comprometerse con reglas de política monetaria, reglas que habrán de ser secundadas por una agencia independiente del Congreso, evitando que este último actúe con total discrecionalidad. Véase Finn Kydland y Edward Prescott, «Rules Rather than Discretion», *Journal of Political Economy* 85 (1977): 473-491.

13. Además de la aceptación de los principios como objetivamente válidos, sería útil disponer de una lista comparativa de las otras bases que resultarían aceptables para confiar en la efectiva realización de las acciones de una persona. Esas bases podrían incluir aquellas otras funciones (enumeradas) de los principios cuyo cumplimiento no depende de una creencia en la validez objetiva del principio.

14. Véase Carol Gilligan, *In a Different Voice* (Cambridge, Mass.: Harvard Univ. Press, 1982); véase también Bill Puka, «The Liberation of Caring: A Different Voice for Gilligan's "Different Voice"», *Hypatia* 5 (1990): 58-82.

15. Siguiendo la tradición filosófica, uso el término *determinar* para significar fijar, causar, provocar la ocurrencia de —como en «determinismo»—, pero obsérvese también el lado estimativo/evidenciativo/epistemológico del término, como cuando decimos «aún no he determinado qué es lo que fulano está tratando de hacer».

16. George Ainslie, «Specious Reward: A Behavioral Theory of Impulsiveness and Impulse Control», *Psychological Bulletin* 82 (1975): 463-496; Ainslie, «Beyond Microeconomics», en *The Multiple Self*, compilado por Jon Elster (Cambridge: Cambridge Univ. Press, 1986), págs. 133-175.

17. ¿Se puede usar la información sobre el grado presente de temporalidad de la preferencia para hacer una estimación aproximada de la hostilidad y del nivel de riesgo entrañados por el medio ambiente y la historia vital del organismo en el que apareció evolutivamente por vez primera esa preferencia temporal? ¿Podríamos usar la información sobre la forma general de la curva de la preferencia temporal para contrastar teorías sobre el dominio en que operó la selección (por ejemplo, cuán amplia es la clase de parentesco en que opera la selección de parentesco)?

18. Analicé por primera vez los peligros del doble descuento en «On Austrian Methodology», *Synthese* 36 (1977): 353-392.

19. Esta última forma es una consecuencia de las ecuaciones de la «ley de ajuste». Véase Richard Herrnstein, «Relative and Absolute Strenghts of Response as a Function of Frequency of Reinforcement», *Journal of the Experimental Analysis of Behavior* 4 (1961): 267-272.

20. Agradezco a Amartya Sen que planteara esta cuestión.

21. También está el fenómeno del *arrepentimiento,* un descenso de la utilidad presente debido a la reconsideración de una acción pasada que ahora se querría haber evitado. Tener una tendencia al arrepentimiento podría ayudarles a ustedes en cierto modo a dejar atrás la tentación en el período B, pues mientras dura ese período ustedes pueden anticipar el descenso del nivel de utilidad en C y posteriormente, si aceptan ahora la recompensa menor y más cercana. La cuestión es si esa anticipación retroalimentará al conjunto de utilidades en B lo suficiente como para afectar a la elección que se hace en ese período.

22. Para una discusión crítica del objetivo único de maximizar la utilidad total a lo largo de un ciclo vital entero, véase mi *The Examined Life* (Nueva York: Simon and Schuster, 1989), págs. 100-102.

23. Véase también Jon Elster, *Ulysses and the Sirens* (Cambridge: Cambridge Univ. Press, 1979).

24. Centrarse en un grupo entero de acciones de un cierto tipo en el ámbito personal traerá a la memoria de algunos lectores el utilitarismo de las reglas en el ámbito público. Lo que aquí nos interesa, en cambio, es la cuestión de hasta qué punto la aceptación de un principio general afecta a la elección de una acción particular que, en ausencia del principio, no acarrearía la máxima utilidad. La cuestión análoga sería hasta qué punto alguien con deseos utilitaristas, que tomara (de algún modo) decisiones de acuerdo con un principio del utilitarismo de las reglas, conseguiría hacer efectivo ese principio en situaciones de elección particulares.

25. La promulgación de un principio afecta también al modo en que terceros lo llevarán a cabo; un diseñador de principios tendrá que tomar en cuenta el modo en que otros puedan distorsionarlos o abusar de ellos. En mi *The Examined Life,* pág. 284, me he referido a una cuestión parecida, el modo en que teóricos sociales como Marx y Freud deberían haber tomado precauciones frente a su vulgarización.

26. Una explicación alternativa de los principios que incorporan universalizadores podría sugerir que los principios codifican razones y que las razones son universales (aun cuando rebatibles); lo que explicaría que los principios lo sean también. Pero ¿por qué las razones no son «para la mayor parte de», sino que son «universales y rebatibles», aun cuando los porcentajes puedan ser los mismos?

27. Con frecuencia, la gente no acepta la doctrina que recomienda ignorar los costes sumergidos, como lo muestran las decisiones que toma cuando se le presentan elecciones hipotéticas. Véase al respecto, H.R. Arkes y C. Blumer, «The Psychology of Sunk Cost», *Organizational Behavior and Human Decision Processes* 35 (1985): 124-140. Arkes y Blumer consideran que la gente que se desvía de la doctrina en el ejemplo de las entradas es irracional.

Scott Brewer se ha preguntado si la persona anticipa con frecuencia que, de no usar la entrada esta noche en particular, le preocupará la posibilidad de comprar otra en el futuro, o de hacer algún gasto recreativo de otro tipo, y el hecho de que dentro de su esquema de contabilidad no cabe el deseo de ese ulterior gasto recreativo. No todos los costes serían, por lo tanto, cosa del pasado. Obsérvese que los economistas condenarían como irracional esa forma de segmentar los esquemas de contabilidad.

28. Véase el ensayo de Bernard Williams en J.J.C. Smart y Bernard Williams, *Utilitarianism: For and Against* (Cambridge: Cambridge Univ. Press, 1973) [trad. cast.: *Utilitarismo*, Madrid, Tecnos, 1981]; Williams, «Persons, Character and Morality», en *The Identities of Persons*, compil. Amelie Rorty (Berkeley: Univ. of California Press, 1976).

29. Debo esta sugerencia a Susan Hurley, quien, en relación con, y paralelamente a, nuestra anterior cuestión acerca de si podemos confiar en alguien que se adhiere a un principio cuando sus únicas razones para mantenerlo son los beneficios que le acarrea a él mismo el que nosotros confiemos en él, se pregunta también si puede alguien esperar respetar los costes que él mismo ha sumergido si él mismo no puede luego pensar que, previamente, tuvo alguna razón independiente para sumergirlos, alguna razón que no fuera la de inducirse a sí mismo a respetarlos.

30. Véase Thomas Schelling, «The Art of Commitment», en su *Arms and Influence* (New Haven: Yale Univ. Press, 1966), págs. 35-91. Véase también la discusión que hace Schelling de la «racionalidad de la irracionalidad».

31. Podría ser también un rasgo útil, sobre todo para los jóvenes, el ser optimistas respecto de las posibilidades de éxito de posibles proyectos —si no, nada nuevo u osado se intentaría nunca—, y sin embargo, tender a atenerse a los proyectos en curso en los que se ha hecho una inversión significativa —si no, a la primera dificultad seria, uno podría dejarlo para empezar otro proyecto, de nuevo con (gran) optimismo—.

32. Una vez que una acción o un resultado viene a simbolizar otras acciones u otros resultados, su presencia puede considerarse como una evidencia para esas otras acciones, o como una causa de ellas, mas ése es un producto de la simbolización, no una parte de su factoría original (aun si ese papel evidencial o causal puede luego reforzar a la conexión simbólica).

33. Eso es lo que presumiría una teoría maximizadora de la decisión. Hay otros tipos de teoría normativa de la decisión, por ejemplo, la teoría «satisfactoria» de Herbert Simon, pero ésta requeriría también que la acción emprendida tuviera, o se le imputara, una utilidad por encima del (cambiante) nivel de aspiración.

34. No es que tenga que ser siempre la expresividad lo que fluye hacia atrás a través de la conexión simbólica. También otras cosas pueden fluir hacia atrás, dando lugar a nuevas características de la acción que tienen, ellas mismas, gran utilidad para el agente. Lo que importa aquí es que no es la utilidad lo que fluye hacia atrás.

35. ¿Deberíamos, pues, distinguir los casos en los que el objetivo es x, y alguien actúa simbólicamente para conseguir x, de los casos en los que el objetivo es una conexión simbólica con x, y alguien actúa instrumentalmente para conseguir esa conexión?

36. Para una discusión de las acciones en equilibrio, véase mi *Philosophical Explanations* (Cambridge, Mass.: Harvard Univ. Press, 1981), págs. 348-352.

37. Agradezco a Bernard Williams la mención de este ejemplo. La última cláusula entre paréntesis, obvio es decirlo, previene la refutación en ausencia de un criterio independiente que permita establecer cuándo algo está suficientemente «ela-

borado». Williams señala también que algunos significados simbólicos entrañan una fantasía que es estrictamente imposible de realizar; y no resulta claro cómo pueden asignarse utilidades a situaciones imposibles. No querría prejuzgar, sin embargo, la posibilidad de que, incluso las situaciones incoherentes, tuvieran una gran utilidad para nosotros.

38. Véase Charles Fried, *An Anatomy of Values* (Cambridge, Mass.: Harvard Univ. Press, 1970), págs. 207-218.

39. Véase Ronald Carson, «The Symbolic Significance of Giving to Eat and Drink», en *By No Extraordinary Means: The Choice to Forgo Life-Sustaining Food and Water*, comp. por Joanne Lynn (Bloomington: Indiana Univ. Press, 1986), págs. 84-88.

40. Véase mi *The Examined Life*, págs. 286-292.

41. Para una discusión del modo en que alguna publicidad de productos comerciales se sirve de este fenómeno, véase mi *The Examined Life*, págs. 121-122.

42. Véase Raymond Firth, *Symbols: Public and Private* (Ithaca: Cornell Univ. Press, 1973); Clifford Geertz, «Deep Play: Notes on the Balinese Cock-Fight», en su *The Interpretation of Cultures* (Nueva York: Basic Books, 1973) [trad. cast.: *La Interpretación de las culturas*, Barcelona, Gedisa, 1988].

43. En realidad, dado el grado de participación de los factores sociales en la creación, mantenimiento, coordinación y contención del significado simbólico, podríamos encontrarnos aquí con una limitación de las explicaciones individualistas metodológicas —una limitación importante, dados los efectos y las consecuencias de esos significados—. Pues una utilidad simbólica podría ser social no sólo por el hecho de estar socialmente moldeada y ser socialmente compartida —tener, esto es, el mismo valor para mucha gente en la sociedad—, sino también por el hecho de ser vista *como* compartida —siendo este último rasgo intrínseco al hecho de tener utilidad simbólica—. No resulta claro cómo las explicaciones individualistas metodológicas habrían de lidiar con las complicaciones que esto conlleva. En cualquier caso, dejando de lado el significado simbólico, no resulta claro cómo podría el individualismo metodológico dar razón del lenguaje.

44. Nelson Goodman, *Languages of Art* (Indianapolis: Bobbs-Merrill, 1968) [trad. cast.: *Los lenguajes del arte*, Barcelona, Seix Barral, 1974], págs. 45-95. Sobre la teoría goodmaniana del mérito estético, véase mi «Goodman, Nelson on Merit, Aesthetic», *Journal of Philosophy* 69 (1972): 783-785.

45. Catherine Elgin, *With Reference to Reference* (Indianapolis: Hackett, 1983), pág. 143, discute una cadena particular con cinco eslabones.

46. ¿Puede la utilidad simbólica de una acción entenderse como una *interpretación* de esa acción, como un modo de verse uno a sí mismo, o de ver la acción de una cierta forma, de manera que los varios mecanismos de eslabonamiento interpretativo, y las mismas teorías globales de la interpretación, pudieran entrar en la definición de la utilidad simbólica?

47. Véase la sección «The Ideal and the Actual» en mi *Examined Life*. Eso abre la posibilidad de que la gente que no quiere que se secunde *P* para obtener determinado resultado pueda arreglárselas de modo que *P* sea secundado para obtener otro resultado monstruoso, consiguiendo así el descrédito del principio.

48. Para algunas reflexiones críticas sobre la doctrina, según la cual somos libres cuando nuestras acciones han sido determinadas autoconscientemente por una ley de la razón, la cual sería un principio constitutivo de nuestra naturaleza esencial, véase mi *Philosophical Explanations*, págs. 353-355.

NOTAS AL CAPÍTULO 2

1. Para una selección de artículos hasta 1985 y un listado bibliográfico de otros, véase *Paradoxes of Rationality and Cooperation: Prisoner's Dilemma and Newcomb's Problem*, comp. Richmond Campbell y Lanning Sowden (Vancouver: University of British Columbia Press, 1985).

2. Sobre la teoría causal de la decisión, véase Allan Gibbard y William Harper, «Counterfactuals and Two Kinds of Expected Utility», en *Foundations and Applications of Decision Theory*, comp. C.A. Hooker y otros (Dordrecht: Reidel, 1978), reproducido en *Paradoxes of Rationality and Cooperation*, comp. Campbell y Sowden; David Lewis, «Causal Decision Theory», *Australasian Journal of Philosophy* 59 (1981): 5-30; J.H. Sobel, «Circumstances and Dominance in a Causal Decision Theory», *Synthese* 63 (1985).

Ni tampoco me apercibí de la posibilidad de situaciones específicas en las que los estados fueran probabilísticamente independientes de las acciones, y sin embargo, estuvieran influidos por ellas —el ejemplo de Reoboam sugerido por Gibbard y Harper—, posibilidad que debería haber figurado como cuarta columna en el cuadro de tres columnas de la pág. 132 de mi artículo.

3. Sobre la maximización de la utilidad condicionalmente esperada, aun sin usar el término *utilidad evidencial*, véase mi tesis doctoral en la Universidad de Princeton de 1963, *The Normative Theory of Individual Choice* (reimpresión; Nueva York: Garland Press, 1990). Véase la pág. 232: «Las probabilidades que han de usarse en la determinación de la utilidad esperada de una acción deben ser ahora probabilidades condicionales de los estados dada la realización de la acción. (Esto es verdad en general. Sin embargo, cuando los estados son probabilísticamente independientes de las acciones, la probabilidad condicional de cada estado dada la realización de una de las acciones será igual a la probabilidad del estado, de manera que habrá que usar esta última.)» También allí se enunciaba para los casos de las dos particulares acciones que se discutían la fórmula para la utilidad condicional esperada, pero no la fórmula general para la acción variable. La fórmula general la de Richard Jeffrey, *The Logic of Decision* (Nueva York: McGraw-Hill, 1965).

Las cuestiones que nos interesan en este libro surgen todas cuando las probabilidades, condicionales o de otro tipo, subjetivas u objetivas, están nítidamente definidas. Otras cuestiones han llevado a otros autores a formular teorías que se sirven de intervalos de probabilidad; véase, por ejemplo, Isaac Levi, *Hard Choices* (Cambridge: Cambridge Univ. Press, 1986). Cómo habría que reformular los puntos de vista presentados aquí para que casaran con esos marcos es una cuestión abierta a la investigación.

4. Los intentos de rechazar el problema como mal construido, mal definido, o imposible en principio, incluyen: Isaac Levi, «Newcomb's Many Problems», *Theory and Decision* 6 (1975): 161-175; J.L. Mackie, «Newcomb's Paradox and the Direction of Causation», *Canadian Journal of Philosophy* 7 (1977): 213-225; y William Talbott, «Standard and Non-Standard Newcomb Problems», *Synthese* 70 (1987): 415-458. Para una defensa del problema frente a varios de esos críticos, véase Jordan Howard Sobel, «Newcomblike Problems», *Midwest Studies in Philosophy* 15 (1990): 224-255.

5. Una excepción es J.H. Sobel, quien en «Infallible Predictors», *Philosophical Review* 97 (1988): 3-24, concluye su artículo considerando «un problema de Newcomb límite», en el que la cantidad en la primera caja se incrementa de 1.000 dólares a (casi) un millón. Sobel, sin embargo, no entra a considerar la posibilidad de reducir los 1.000 dólares de la primera caja a casi nada. En Kenneth MacCrimmon y Stig Larsson, «Utility Theory: Axioms versus "Paradoxes"», en *Expected Utility Hipothe-*

sis and the Allais Paradox, comp. Maurice Allais y Ole Hagen (Dordrecht: Reidel, 1979), pág. 393, se consideran las consecuencias de variar la cantidad de la segunda caja, pero no la de la primera.

6. Si no llegar a tener confianza completa en un principio lleva a una persona a secundar una combinación de principios, ¿qué ocurre si no llega a tener completa confianza en esa combinación? Si hay algún otro determinado principio en el que uno tiene *alguna* confianza, entonces, en la medida en que el argumento depende sólo de los grados reales de confianza, parece que este otro principio debería incluirse también en la ponderación.

7. Para una noción diferente de las consideraciones evidencialistas, según la cual éstas sólo resultan atractivas cuando encajan con el razonamiento cooperativo en situaciones interpersonales, véase Susan Hurley, «Newcomb's Problem, Prisoners' Dilemma, and Collective Action», *Synthese* 86 (1991): 173-196.

8. «Pero ¿qué es lo que explica el desacuerdo entre los propugnadores de *UCE* y los propugnadores de *UEE*? ¿Es un desacuerdo fáctico o un desacuerdo axiológico?» Este interrogante supone que ambos comparten una fórmula UE y pregunta si su desacuerdo radica en el componente de probabilidad o en el de utilidad. Sin embargo, si la fórmula *VD* es correcta, entonces hay *otras* cosas sobre las que discrepar, incluidos los pesos *Wc* y *We*, la naturaleza de la fórmula y también —anticipándonos a los próximos párrafos— la incorporación de otros factores. Preguntar «¿hecho o valor?» —sin permitir otra alternativa— es presumir que lo que *debe* haber en común es el simple marco *UE*, y que sólo *en* ese marco puede surgir el desacuerdo.

9. David Gauthier considera la cuestión de qué disposición de elección debería elegir tener una persona en *Morals by Agreement* (Oxford: Oxford Univ. Press, 1985), cap. 6, secs. 2-3.

10. Nozick, «Newcomb's Problem», pág. 125; Gibbard y Harper, «Counterfactuals».

11. Rudolf Carnap sostuvo (*The Logical Foundations of Probability* [Chicago: Univ. of Chicago Press, 1950]) que las sentencias que afirman que «el grado de confirmación de h basado en e es n» son, si verdaderas, analíticas. Sin embargo, incluso él sostuvo también que la cuestión de *qué* particular función de confirmación hay que elegir (c^*, c^+, o cualquiera que se halle en el *continuum* de los métodos inductivos), y por lo tanto, cuál definirá esa relación analítica, es una cuestión de elección pragmática y dependerá de los hechos generales acerca del universo.

12. Véase John Milnor, «Games against Nature», en *Decision Processes*, comp. R.M. Thrall, C.H. Coombs y R.L. Davis (Nueva York: John Wiley, 1954), págs. 49-59, y R.D. Luce y Howard Raiffa, *Games and Decisions* (Nueva York: John Wiley, 1957), págs. 275-298. Ya dije antes que no es necesario que el significado simbólico traslade la proporcionalidad a los contextos probabilísticos. No obstante, la fórmula *VD* incluye la utilidad simbólica como uno de los componentes ponderados. Podríamos sorprendernos entonces de que la utilidad simbólica se traslade al contexto ponderado *VD*. Ello es, sin embargo, que un traslado a una situación probabilística es un traslado a una situación *diferente*, mientras que un traslado a la fórmula *VD* no altera la situación de elección.

13. El único estudio psicológico que conozco al respecto aborda tanto las conexiones causales, cuanto las evidenciales, y trata de desenredarlas y separarlas. Véase G.A. Quattrone y Amos Tversky, «Causal versus Diagnostic Contingencies: On Self-Deception and the Voter's Illusion», *Journal of Personality and Social Psychology* 46 (1984): 237-248.

14. No todo modo de acción entraña una conexión con una consecuencia, una conexión incorporable a una fórmula en tanto que conexión causal, evidencial o sim-

bólica. Considérese la actuación sin motivos, un modo de acción cuyas variantes aparecen en la literatura del budismo, del taoísmo y del hinduismo. En esa literatura, la persona no actúa para llegar a ser de algún modo, o para ser de algún modo, o para producir resultados, o para adquirir evidencias, o para simbolizar algo. Acaso actúe para *alinearse* a sí misma (correctamente) con la realidad más profunda, para ser alineada con esa realidad dejando que ésta actúe a través de ella. Este modo de acción requiere más análisis, pero no parece entrañar un modo de conexión con una consecuencia.

15. H.P. Grice, «Meaning», *Philosophical Review* 67 (1957): 377-388.

16. Véase John von Neumann y Oskar Morgenstern, *The Theory of Games and Economic Behavior,* 2.ª ed. (Princeton: Princeton Univ. Press, 1947), apéndice. Una exploración de las cuestiones filosóficas planteadas por el conjunto de condiciones de Von Neumann-Morgenstern y otros conjuntos parecidos puede encontrarse en Robert Nozick, *The Normative Theory of Individual Choice* (Tesis doctoral, Princeton University, 1963; reimpresión, Nueva York: Garland Press, 1990).

17. Así trata L.J. Savage a las acciones dentro del formalismo de su teoría de la decisión; véase su *The Foundations of Statistics* (Nueva York: John Wiley, 1954). Sin embargo, una acción no es susceptible de tamaña reducción, aun dejando de lado las cuestiones atinentes a su posible valor simbólico. Véase mi *The Normative Theory of Individual Choice,* págs. 184-193.

18. Véase, por ejemplo, Peter Hammond, «Consequentialist Foundations for Expected Utility», *Theory and Decision* 25 (1988): 25-78.

19. A resultas del problema de Newcomb se han investigado casos en los que la *probabilidad* de un resultado se ve alterada por las razones para realizar la acción, lo que ha dado lugar a la bibliografía sobre «ratificabilidad».

20. Vale la pena mencionar también que, cuando la secuencia de las acciones es estratégicamente relevante, los cultivadores de la teoría de los juegos no se concentran simplemente en la representación matricial del juego y en sus beneficios, sino que atienden también a la representación extensiva del árbol del juego.

21. Gibbard y Harper, «Counterfactuals and Two Kinds of Expected Utility», pág. 151.

22. David P. Kreps, P. Milgrom, J. Roberts y R. Wilson, «Rational Cooperation in the Finitely Repeated Prisoner's Dilemma», *Journal of Economic Theory* 27 (1982): 245-252.

23. Como ha dicho sumariamente un autor, un jugador podría «elegir una acción fuera de equilibrio para generar en otros jugadores creencias y estrategias fuera de equilibrio». Eric Rasmussen, *Games and Information* (Oxford: Basil Blackwell, 1989), pág. 111.

24. Una nota marginal de pasada. En mi tesis doctoral de 1963 vi la necesidad, en situaciones de teoría de juegos, de niveles de conocimiento infinitamente extendidos, en los que cada jugador conociera la estructura de la situación del juego tal como lo concibe la teoría, y en la que cada uno conociera que los demás conocen que él conoce que los demás conocen, etc. (*The Normative Theory of Individual Choice,* pág. 274). Pero yo creía que esto era el chocolate del loro. Poco imaginaba yo entonces el interés y las vastas implicaciones que la condición del conocimiento común de la racionalidad iban a traer consigo. Véase Robert Aumann, «Correlated Equilibrium as an Expression of Bayesian Rationality», *Econometrica* 55 (1987): 1-18, y Drew Fudenberg y Jean Tirole, *Game Theory* (Cambridge, Mass.: M.I.T. Press, 1991), págs. 541-572.

25. Sobre la estrategia de toma-y-daca, véase Robert Axelrod, «The Emergence of Cooperation among Egoists», reproducido en *Paradoxes of Rationality and Coo-*

peration, comp. Campbell y Sowden, así como su *The Evolution of Cooperation* (Nueva York: Basic Books, 1984).

26. Hablo aquí de un modo intuitivo, pues en una medida de escala de intervalo con un punto cero arbitrario, no reviste especial importancia el que una cantidad medida sea negativa. Una unidad arbitraria y un punto cero arbitrario son una fuente de problemas para las comparaciones interpersonales de utilidad. Para algunas propuestas al respecto, véase mi «Interpersonal Utility Theory», *Social Choice and Welfare* 2 (1985): 161-179.

27. Véase Philippa Foot, «The Problem of Abortion and the Doctrine of the Double Effect», en su *Virtues and Vices* (Berkeley: Univ. of California Press, 1978), págs. 19-32; Judith Thompson, «Killing, Letting Die, and the Trolley Problem» y «The Trolley Problem», en su *Rights, Restitution and Risk* (Cambridge, Mass.: Harvard Univ. Press, 1986), págs. 78-116; Warren Quinn, «Actions, Intentions, and Consequences: The Doctrine of Double-Effect», *Philosophy and Public Affairs* 18 (1989): 334-351; Warren Quinn, «Actions, Intentions, and Consequences: The Doctrine of Doing and Allowing», *Philosophical Review* 98 (1989): 287-312; Frances Kamm, «Harming Some to Save Others», *Philosophical Studies* 57 (1989): 227-260.

28. A otros efectos, podríamos querer extender esa matriz, añadiéndole una tercera dimensión que representara la magnitud de la consecuencia. (Justin Hughes me sugirió extender de esa forma la matriz en contextos jurídicos. En tales contextos podríamos desear saber cuán mala es la consecuencia: cuán mala era la consecuencia que se pretendía, cuán mala fue la que ocurrió.) Dentro de la teoría de la decisión, sin embargo, va de suyo que esta magnitud está ya representada por la utilidad del resultado.

29. Recuérdese también la discusión que hemos hecho antes del modo en que los metaprincipios que ordenan no violar ningún principio podrían hacer que una violación valiera por todas, confiriendo así a todo principio un fuerte peso deontológico.

30. Véase Amartya Sen, *Ethics and Economics* (Oxford: Basil Blackwell, 1987) [trad. cast.: *Sobre ética y economía*, Madrid, Alianza, 1989], págs. 80-88.

NOTAS AL CAPÍTULO 3

1. ¿Qué si el *proceso* más efectivo para alcanzar el objetivo de realizar la mejor acción no entraña nada que podamos llamar un «procedimiento», nada que tenga que ver con el control consciente de medidas o la consciente aplicación de reglas, principios o razones de alguna clase?

2. Para diversos enfoques de este asunto, véase Isaac Levi, *The Enterprise of Knowledge* (Cambridge, Mass.: M.I.T. Press, 1980); Gilbert Harman, *Change in View* (Cambridge, Mass.: M.I.T. Press, 1986); y Peter Gardenfors, *Knowledge in Flux* (Cambridge, Mass.: M.I.T. Press, 1988).

3. Véase Alvin Goldman, *Epistemology and Cognition* (Cambridge, Mass.: Harvard Univ. Press, 1986), págs. 58-121, quien ofrece un análisis de la justificación, más que de la racionalidad, en términos de fiabilidad; Frank Ramsey, «Reasonable Degrees of Belief», en su *The Foundations of Mathematics and Other Logical Essays* (Londres: Routledge and Kegan Paul, 1931), págs. 199-203; William Talbott, *The Reliability of the Cognitive Mechanism* (Tesis Doctoral, Harvard Univ., 1976; reimp. con un nuevo prólogo Nueva York: Garland Press, 1990); Stephen Stich, *The Fragmentation of Reason* (Cambridge, Mass.: M.I.T. Press, 1990), págs. 89-100.

4. Véase Thomas Kuhn, *The Essential Tension* (Chicago: Univ. of Chicago Press,

1977) [trad. cast.: *La tensión esencial*, Madrid, FCE, 1983], págs. 320-339; W.V. Quine y Joseph Ullian, *The Web of Belief*, 2.ª ed. (Nueva York: Random House, 1978), págs. 64-82.

5. John Rawls, *A Theory of Justice* (Cambridge, Mass.: Harvard Univ. Press, 1971) [trad. cast.: *Teoría de la justicia*, Madrid, FCE, 1979], págs. 62, 90-95.

6. Hay situaciones de teoría de juegos en las que la ignorancia de una probabilidad correcta puede resultar beneficiosa. Véase Eric Rasmussen, *Games and Information* (Oxford: Basil Blackwell, 1989), pág. 116, «Entry Deterrence IV».

7. Compárese la distinción entre objetivos y restricciones laterales con la discusión de un «utilitarismo de los derechos» que se hace en mi *Anarchy, State, and Utopia* (Nueva York: Basic Books, 1974), págs. 28-33.

8. Amartya Sen, «Rights and Agency», *Philosophy and Public Affairs* 11 (1982).

9. La literatura convierte a la figura paterna en mujer y al hijo en varón. ¿Son más amantes las madres? (Los hijos varones *son* más prontos al crimen.) ¿O presume la literatura que las mujeres son más susceptibles de albergar conflictos entre la emoción y la evidencia?

10. Éste, como hemos visto, es el criterio de la mejor acción. Supongamos, pues, que ella llega a su creencia a través de un proceso que fiablemente produce las mejores acciones.

11. Sobre la «ética de la creencia», véase William James, «The Will to Believe», en su *The Will to Believe and Other Essays* (Cambridge, Mass: Harvard Univ. Press, 1979), págs. 13-33; Jack Meiland, «What Ought We to Believe», *American Philosophical Quarterly* 17 (1980): 15-24; John Heil, «Believing What One Ought», *Journal of Philosophy* 80 (1983): 752-765. Heil hace una distinción parecida a la mía entre la proposición de que *p* es lo que racionalmente hay que creer y la creencia de que *p* es lo que hay que hacer racionalmente.

12. Véase Frederick Schauer, *Playing by the Rules* (Oxford: Clarendon Press, 1971), que contiene una amplia discusión de esta cuestión en relación con las reglas. ¿Tienen éstas una autoridad con peso propio, aun en casos particulares en los que los propósitos últimos para cuya promoción estaban pensadas no resultan promovidos, o resultan incluso claramente obstaculizados por ellas?

13. Los teóricos tradicionales de la racionalidad se concentran en razones y en el razonamiento, sin mencionar los procesos fiables; algunos teóricos más recientes se han centrado en el aspecto de fiabilidad sin poner de relieve el razonamiento. Tales exclusiones resultan comprensibles si ciertos tipos de razonamiento y de fiabilidad van siempre de consuno, si los únicos procesos fiables tienen que ver con tipos de razonamiento —los golpes en la cabeza no funcionan— y esos tipos de razonamiento son siempre fiables.

14. Véase Karl Popper, *The Logic of Scientific Discovery* (Nueva York: Basic Books, 1959) [trad. cast.: *La lógica de la investigación científica*, Madrid, Tecnos, 1985], y *Conjectures and Refutations* (Nueva York: Basic Books, 1962) [trad. cast.: *Conjeturas y refutaciones*, Barcelona, Paidós, ^{3}1991], pág. 240.

15. Podríamos sugerir aquí una exigencia comparable a la noción de rastreo de la pista en la teoría del conocimiento. Véase mi *Philosophical Explanations* (Cambridge, Mass: Harvard Univ. Press, 1981), cap. 3.

16. Véase Robert Nozick, «Moral Complications and Moral Structures», *Natural Law Forum* 13 (1968): 1-50; *Readings in Nonmonotonic Reasoning*, comp. Matthew Ginsberg (Los Altos, Calif.: Morgan Kaufmann, 1987); John Pollock, *How to Build a Person* (Cambridge, Mass.: M.I.T. Press, 1989), págs. 124-155.

17. Tres personas han sido juzgadas por asesinato y halladas culpables, pero sólo una ha sido sentenciada a la pena capital. Ninguno de ellos sabe cuál, y podemos

presumir, como cada uno de ellos lo hace, que cada uno tiene una probabilidad de 1/3 de ser ejecutado. La víspera de la ejecución, el prisionero A pide al guardián, quien sabe qué prisionero va a ser ejecutado, que entregue una nota que él ha redactado para su mujer a uno de los otros dos, al que no vaya a ser ejecutado. Cuando el guardián recibe la nota, el prisionero A cree que tiene una probabilidad de 1/3 de ser ejecutado al amanecer. Cuando el guardián regresa y le comunica verazmente que ha entregado la nota, el prisionero A aún cree que tiene una probabilidad de 1/3 de ser ejecutado. No ha recibido ninguna información nueva relevante, pues él ya sabía de antemano que (al menos) uno de los otros dos prisioneros no sería ejecutado y, por lo tanto, estaría en condiciones de entregar la nota a su mujer. Entonces le pregunta al guardián a qué prisionero ha entregado la nota, y el guardián le responde que al prisionero B. Erraría el prisionero A si razonara que ahora tienen una probabilidad de 1/2 de ser ejecutado basándose en la idea de que él y el prisionero C empezaron con una idéntica probabilidad igual a 1/3 y que la situación se mantiene simétrica, de manera que ahora siguen teniendo una probabilidad idéntica, que ahora a subido hasta 1/2. Esta particular situación no es simétrica. Tanto B como C hubieran podido ser los destinatarios de la nota (de manera que el que C no la reciba *es* relevante para estimar la probabilidad de su ejecución —que ahora sube hasta 2/3—), mientras que A no podría haber sido el destinatario de la nota. Podría haber llegado la información de que C no iba a ser ejecutado, pero no podría haber llegado la información de que A no iba a ser ejecutado. De aquí que, cuando llega la información real, lo que se ha incrementado es la probabilidad de C. En cambio, la probabilidad de que el prisionero A sea ejecutado se *incrementaría* hasta 1/2 en las situaciones siguientes: el prisionero A pregunta al guardián cuando éste regresa: «¿Recibió el prisionero B la nota, sí o no?, y el guardián contesta «sí»; o el prisionero A le pide al comienzo al guardián que entregue la nota a un prisionero que no será ejecutado el próximo día, a cualquiera de los *tres*, y (suponiendo que el guardián puede entregar la nota con igual probabilidad a cualquier prisionero que no vaya a ser ejecutado, incluido A) el guardián regresa diciendo que la entregó al prisionero B. El factor crucial es la manera en que difiera la recepción de la nota por parte del prisionero B, según vaya a ser ejecutado el prisionero A o el prisionero C. Si A va ser ejecutado, la probabilidad de que B reciba la nota es de 1/2 (y la probabilidad de que la reciba C es la misma). Si C va a ser ejecutado, la probabilidad de que la reciba B es 1. Esos valores de probabilidad suman 1 y 1/2, de los cuales la probabilidad que procede de que A sea ejecutado (es decir, 1/2) constituye un tercio de este total. De aquí que la información de que B recibió la nota deje a A con una probabilidad de ser ejecutado de 1/3 y conceda a C una probabilidad de 2/3. Todo esto resulta transparente en un diagrama reticular bayesiano; véase Judea Pearl, *Probabilistic Reasoning in Intelligent Systems: Netwoks of Plausible Inference* (San Mateo, Calif.: Morgan Kaufmann, 1989), fig. 9.1, pág. 417.

18. Lo que importa no es sólo el particular contenido de *e*, sino también qué otra información podría haber llegado (y con qué probabilidades). Haber tratado de inyectar esta última información en la evidencia misma habría alterado radicalmente la estructura de la lógica inductiva carnapiana. Véase Rudolf Carnap, *The Logical Foundations of Probability* (Chicago: Univ. of Chicago Press, 1950), y *The Continuum of Inductive Methods* (Chicago: Univ. of. Chicago Press, 1952). El sistema preferido por Carnap en el primero de estos libros incluía una elaborada versión del principio de indiferencia: todas las descripciones de estructura reciben la misma probabilidad *a priori*. Quizás al nivel más básico deba haber una razón estructural para cualquier diferencia en la probabilidad *a priori*. (Aun en este caso, podría pensarse

que esto es una cuestión de apoyos empíricos.) Mas no hay ninguna razón para pensar que hemos llegado a este nivel, ni siquiera en la física fundamental de nuestros días, y desde luego no en las propiedades que consideramos de ordinario.

19. Consideren una pieza *I* de información procedente de sus fuentes de información y que dice que «*p* es verdadera». ¿Qué revela la recepción de *I*? Aplicando el teorema de Bayes, prob(*p*/*I* se reciba) = [prob(*I* se reciba/*p*) × prob inicial(*p*)]/[prob(*I* se reciba/*p*) × prob inicial(*p*)] + [prob(*I* se reciba/no-*p*) × prob inicial(no-*p*)]. Supongamos adicionalmente que las fuentes de información dirán o bien que *p* es verdadera, o bien que no lo es. (Considerar la posibilidad de que no digan nada sería complicar innecesariamente la formulación a seguir.) Entonces prob(*I* se reciba/no-*p*) = 1 − prob(fuentes digan «*p* es falso»/no-*p*). De modo que el denominador en la formulación anterior del teorema de Bayes equivale a [prob(*I* se reciba/*p*) × prob inicial(*p*)] + [1 − prob(fuentes digan «*p* es falso»/no-*p*) × prob inicial(no-*p*)]. La última parte de este denominador, después del signo más, es igual a la prob inicial(no-*p*) − prob(fuentes digan «*p* es falso»/no-*p*) × prob inicial(no-*p*). De manera que cuanto menor la probabilidad de que estas fuentes digan «*p* es falso», dado no-*p*, tanto menos apoyo dará a la verdad de la hipótesis la recepción de la información *I* diciendo que *p* es verdadero. El teorema de Bayes nos enseña que también debemos considerar qué otra información podría habernos llegado y con qué probabilidades (condicionales).

Atendamos ahora al análisis bayesiano del modo en que la hipótesis epistemológica del escéptico, *SK*, se comporta frente a nuestra observación y experiencia, *E*. Puesto que la hipótesis del escéptico *SK* ha sido armada de manera que la prob(*E*/*SK*) = 1, de ello se sigue que, aun si la prob(*E*/no*SK*) sea también 1, la prob(no-*SK*/*E*) no crecerá por encima de la probabilidad previa de no-*SK*. En un enfoque bayesiano de la probabilidad posterior, la probabilidad posterior del no escéptico es mayor que *su* probabilidad previa. La evidencia no sirve de ayuda.

20. La bibliografía sobre el «equilibrio reflexivo» entre principios y casos aparte del supuesto de que los principios mismos tienen una autoridad independiente y actual en su propio tenor literal. Véase Nelson Goodman, *Fact, Fiction, and Forecast* (Cambridge, Mass: Harvard Univ. Press, 1955), cap. 4, sec. 2, págs. 62-66, y Rawls, *A Theory of Justice* [trad. cast.: *Teoría de la justicia*, Madrid, FCE, 1979], págs. 19-21, 48-51. Para una discusión crítica del equilibrio reflexivo, véase Stich, *The Fragmentation of reason*, págs. 83-89.

21. Para el marco de una teoría de este tipo, véase el fascinante libro de John Holland, Keith Holyoak, Richard Nisbett y Paul Thagard, *Induction: Processes of Inference, Learning, and Discovery* (Cambridge, Mass.: M.I.T. Press, 1986).

22. Véase *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, comp. James McClelland y David Rumelhart, 2 vols. (Cambridge, Mass.: M.I.T. Press, 1986), esp. caps. 1-8, 11, 14, 26.

23. Las simulaciones son ahora mucho más corrientes en las ciencias físicas y sociales, pero los filósofos de la ciencia no se han interesado todavía, por lo que sé, en las cuestiones especiales que se le plantean a la teoría de la explicación cuando una ciencia produce no un cuerpo de conocimientos teóricos y de leyes generales, sino un programa y una simulación.

24. Véase John Holland, *Adaptation in Natural and Artificial Systems* (1975; reimp. Cambridge, Mass.: M.I.T. Press, 1992), págs. 176-179; Holland, Holyoak, Nisbett y Thagard, *Induction*, págs. 70-75, 116-117.

25. Sobre el aislamiento de las contradicciones del conjunto de paradojas teóricas y semánticas, así como sobre la limitación de los daños que producen, véase Ludwig Wittgenstein, *Remarks on the Foundations of Mathematics* (Oxford: Basil Black-

well, 1956) [trad. cast.: *Observaciones sobre los fundamentos de la matemática*, Madrid, Alianza, 1987], II 80-82, III 60, V 8-12.

26. Véase Paul Churchland, *A Neurocomputational Perspective* (Cambridge, Mass.: M.I.T. Press, 1989); Andy Clark, *Microcognition: Philosophy, Cognitive Science, and Parallel Distributed Processing* (Cambridge, Mass.: M.I.T. Press, 1989); Patricia Churchland y Terrence Sejnowski, *The Computational Brain* (Cambridge, Mass.: M.I.T. Press, 1992).

27. En «Simplicity as Fallout» (en *How Many Questions: Essays in Honour of Sidney Morgenbesser*, comp. Leigh Cauman [Indianapolis: Hackett, 1983]), presenté una conjetura sobre cómo la satisfacción de una máxima de simplicidad podría surgir del funcionamiento de un sistema más que ser un componente evaluativo del mismo. El sistema allí considerado no contenía retroalimentación de acuerdo con alguna regla de corrección de errores.

28. El estudio del alcance y los límites de los procedimientos de aprendizaje efectivo ha experimentado recientemente un desarrollo formal generando muchas distinciones útiles (para cada enunciado en un ámbito, ¿hay un momento en el que el procedimiento llegará a la verdad acerca del enunciado? ¿Hay un momento para cada enunciado en un ámbito, tal que el procedimiento llegará a la verdad acerca de todos esos enunciados?) y muchos resultados intrigantes. Véase Daniel Osherson, Michael Stob y Scott Weinstein, *Systems That Learn* (Cambridge, Mass.: M.I.T. Press, 1986).

29. Obsérvese que esta medida evita el «problema de la evidencia conocida». Véase Clark Glymour, *Theory and Evidence* (Princeton: Princeton Univ. Press, 1980), págs. 85-93; Daniel Garber, «Old Evidence and Logical Omniscience in Bayesian Confirmation Theory», en *Testing Scientific Theories*, comp. John Earman (Minneapolis: Univ. of Minnesota Press, 1983), págs. 99-131; Colin Howson y Peter Urbach, *Scientific Reasoning: The Bayesian Approach* (La Salle, Ill.: Open Court, 1989), págs. 270-275; John Earman, *Bayes or Bust* (Cambridge, Mass.: M.I.T. Press, 1992), cap. 5. Pues, aun si *e* es ya conocido y la probabilidad condicional de *e* en *h*1 es *1*, no necesariamente es igual a 1 la prob$(h1 \rightarrow e)$, la probabilidad de que si *h*1 fuera verdadera *daría lugar* a *e*. Si es igual a 1, entonces *h*1 ganará en valor a lo largo de *esa* medida consiguiendo que su probabilidad entre en el numerador.

30. Véase Gilbert Harman, «The Inference to the Best Explanation», *Philosophical Review* 70 (1965): 88-95; Norwood Russell Hanson, *Patterns of Discovery* (Cambridge: Cambridge Univ. Press, 1958), págs. 85-92.

31. Quizá deberíamos interpretar que los propugnadores de la inferencia hacia la mejor explicación la proponen como un principio rebatible de inferencia.

32. Para estimar el grado de apoyo explicativo para una hipótesis de un conjunto de hechos, ¿deberíamos tomar simplemente la conjunción de esos hechos y tratarla como evidencia *e* en nuestra fórmula bayesiana causalizada, o, cuando esos hechos fueran lógicamente independientes, deberíamos tomarlos individualmente, estimar su medida bayesiana causalizada uno a uno y luego sumar esos valores?

33. Para la teoría de una retícula de probabilidades bayesianas condicionales, véase Pearl, *Probabilistic Reasoning in Intelligent Systems*.

34. Compárese con el «método de la tenacidad» de Charles Peirce («The Fixation of Belief», en *The Philosophy of Charles Peirce*, comp. Justus Buchler [Londres: Routledge and Kegan Paul, 1940], págs. 5-22), con la noción de «atrincheramiento» de Nelson Goodman (*Fact, Fiction, and Forecast*, págs. 87-120, y Robert Schwartz, Israel Scheffler y Nelson Goodman, «An Improvement in the Theory of Projectability», *Journal of Philosophy* 67 [1970]: 605-608), y con el sistema de licitación en Holland, Holyoak, Nisbett y Thagard, *Induction*, págs. 70-78, 116-121.

35. Véase Holland, Holyoak, Nisbett y Thagard, *Induction*, pág. 9.

36. Podríamos entender al escéptico epistemológico como si estuviera proponiendo una regla más estricta; no creas un enunciado si su valor de credibilidad es menor que el que podría llegar a tener. Esa regla puede tener diferentes versiones: si el valor de credibilidad del enunciado es menor que el de cualquier otro enunciado, sea o no ese enunciado incompatible con el primero; si su valor de credibilidad fuera aumentado por alguna otra evidencia o razón; si es lógicamente posible que algún enunciado tenga un valor de credibilidad mayor.

37. El teorema de Bayes causalizado estima el grado de apoyo explicativo, para una hipótesis considerando también, en el denominador, todas las hipótesis alternativas. El resultado entra entonces como un factor más en el valor de credibilidad de esa hipótesis. La regla 1 nos exige comparar la credibilidad de la hipótesis no con todas las hipótesis incompatibles, sino con la hipótesis incompatible más creíble. Si y sólo si una hipótesis pasa este test, sobrevivirá como una candidata para la creencia.

38. Si la ulterior investigación muestra que esta prevención hacia la regla 2' carece de fundamento, este requisito adicional sería innecesario.

39. Compárese con el peligro que corre el utilitarismo de las reglas de colapsarse y quedar reducido al utilitarismo de las acciones.

40. Véase Henry Kyburg Jr., *Probability and the Logic of Rational Belief* (Middletown, Conn.: Wesleyan Univ. Press, 1961), págs. 196-199, y «Conjunctivitis», en *Induction, Acceptance, and Rational Belief*, comp. Marshall Swain (Dordrecht: Reidel, 1970), págs. 55-82.

41. Véase Richard Foley, «Evidence and Reasons for Belief», *Analysis* 51, n. 2 (1991): 98-102; Richard Jeffrey, «The Logic of Decision Defended», *Synthese* 48 (1981): 473-492.

42. Algunas investigaciones psicológicas apuntan a que ciertos estilos optimistas de explicar los acontecimientos personalmente negativos —atribuyéndolos a factores momentáneos, delimitados, externos— tienen mejores consecuencias personales que otros modos de explicación —que los atribuyen a factores permanentes externos de carácter general—. A pesar de esas consecuencias personales beneficiosas para el éxito profesional, para la felicidad y quizá para la salud física, es posible que la gente que posee un estilo explicativo pesimista tenga una visión más exacta del mundo. Véase Martin Seligman, *Learned Optimism* (Nueva York: Pocket Books, 1992); sobre la cuestión de la exactitud de la visión del mundo, véanse págs. 108-112 y las referencias citadas en la pág. 298. Esas creencias menos exactas nos acercan a la imagen de personas que creen enunciados con un bajo valor de credibilidad —eso dependería de los pesos que su mecanismo de procesamiento confiera y del modo en que éstos surjan— con (aunque no necesariamente para conseguir) un efecto beneficioso.

43. Para un examen de este asunto y de otros relacionados con él, véase Daniel Dennett, *Consciousness Explained* (Boston: Little, Brown, 1991) [trad. cast.: *La conciencia explicada*, Barcelona, Paidós, 1995], págs. 173-182.

44. Para una discusión de este punto, véase mi *Philosophical Explanations*, págs. 703-706.

45. El término «bayesianismo radical» fue acuñado por Richard Jeffrey. Véase su *Probability and the Art of Judgment* (Cambridge: Cambridge Univ. Press, 1992), ensayos 1 y 4-6.

46. Véase Isaac Levi, *The Enterprise of Knowledge*, págs. 2-19 y *passim*, y *The Fixation of Belief and Its Undoing* (Cambridge: Cambridge Univ. Press, 1991), págs. 57-62 y *passim*.

47. Suspender una creencia *p* presentamente mantenida, y por lo tanto, de la que

piensan que no hay posibilidades serias de que esté equivocada, les abre a ustedes la posibilidad de llegar a adoptar posteriormente una creencia *q* incompatible con *p*. (Esa creencia *q* podría incluso llegar a ser no-*p*.) Ahora, ustedes piensan que, hacerlo, sería sin duda un error. ¿No debería entonces una persona negarse a suspender una creencia presente (en situaciones en las que sus creencias aún no han caído en la contradicción)? Levi maneja este problema como sigue (véase Isaac Levi, *The Fixation of Belief and Its Undoing*, págs. 160-164). Suspender una creencia («contracción») no puede por sí solo llevarles a ustedes a ningún error que no hayan cometido ya, porque no se añade información nueva alguna. Sin embargo, suspender una creencia les pone a ustedes en disposición de añadir una creencia errónea en la etapa ulterior («expansión»), porque esta nueva creencia ya no es incompatible con lo que ustedes creen. Levi lidia con esta dificultad diciendo que sólo deberíamos atender en cualquier momento dado a los resultados de nuestro próximo movimiento o de nuestra próxima decisión, no adónde podríamos ir a parar en el límite infinito de la investigación —él llama a esta última preocupación «mesiánica»—. Pero hay muchas cosas que caen entre la próxima etapa y la plenitud de los tiempos, en particular, la etapa inmediatamente siguiente a la próxima. Es *extremadamente* implausible que no debamos prestarle atención *en absoluto*, pero Levi se ve forzado a ello por su idea de que creer algo entraña tratarlo y usarlo como un criterio de posibilidad seria en todos y cada uno de los contextos, esto es, tratarlo así mientras se siga albergando la creencia —lo que hace verdaderamente difícil suspender una creencia, empujando así a Levi al expediente desesperado de entregarse a una miopía radical—.

48. Véase John W. Payne, James Bettman y Eric Johnson, «The Adaptive Decision Maker: Effort and Accuracy in Choice», en *Insights in Decision Making*, comp. Robin Hogarth (Chicago: Univ. of Chicago Press, 1990), págs. 129-153.

49. Peirce sostiene que en cada contexto hay algo que no resulta dudoso, algo que se toma como dado y excluye otras posibilidades; sin embargo, no es necesario que haya nada en particular que se tome como dado en todos los contextos. Levi sostiene que, en cualquier momento dado, cualquier cosa que sea una creencia sea tomada como dada en todos los contextos (aunque algunos contextos podrían llevarnos a revisar algunas de estas creencias). Esto, como queda dicho, es demasiado fuerte. El proyecto cartesiano era aún más fuerte: hallar algunas creencias que podrían ser tomadas como dadas en todos los contextos y que nunca habrían de necesitar revisión. Se podría argumentar que es permisible tomar *q*1 como dada en el contexto C1 por la vía de hallar algún contexto C2 en el que *q*2 se tomara como dada y de concluir en C2 que *q*1 puede ser tomado como dado en C1. Para que este argumento tenga peso, *q*2 debe ser más débil que *q*1 (y análogamente, C2, más débil o más abstracto). Siguiendo hacia atrás este proceso, podría esperar alcanzarse un contexto en el que nada se tomara como dado —la situación cartesiana de duda radical— y en el que, sin embargo, pudiera justificarse algo, de manera que a partir de entonces pudiera tomarse siempre ese algo como dado. Muchos autores han objetado que Descartes toma como dada la fiabilidad de su razonamiento en esta misma situación de duda radical. También podemos preguntarnos si toma como dado el criterio que fija lo que él puede inferir en esta situación. Parece que el criterio de Descartes es aquí el siguiente: *p* puede aceptarse sin dudas si un demonio malevolente no pudiera convencerme de *p* cuando *p* es falso. Pero este criterio resulta satisfecho por «Un demonio me está engañando» o «Un demonio está actuando sobre mí». Sin embargo está fuera de disputa que no podemos creer y tomar de aquí en adelante como dado este enunciado. (De modo que este criterio es, a lo sumo, una condición necesaria para la creencia cierta.) Podría formularse un criterio más ade-

cuado para establecer la legitimidad de la creencia, pero también él dejaría el flanco abierto a contraejemplos y dificultades. Parece que Descartes no sólo debe de argumentar correcta y fiablemente que ha sido satisfecho un criterio particular; también debe argumentar correcta y fiablemente que este particular criterio es adecuado.

50. Así como la creencia excluye alternativas, así también las excluye el objetivo. ¿Cómo, pues, difiere una creencia de un objetivo (que fija una preferencia estructurada o utilidad), dado que la conducta que elijo es una función de ambos? La conducta B que elijo es una función de mi objetivo g y de mi creencia de que (probablemente) B conseguirá g.

$$B = f(g, \mathit{creen}[\mathrm{prob}(g/B) = m]),$$

en donde m es alto. Esta creencia acerca de la probabilidad de g, a su vez, es alguna función f' de mis otras creencias *creen* (en la medida en que tienen que ver con distintos modos de conseguir el objetivo g). Substituyendo, tenemos

$$B = f(g, f'[\mathit{creen}]).$$

Podemos suponer que la función f tiene que ver con algo así como la fórmula de la utilidad esperada; la función f' tiene que ver con alguna fórmula acerca de la determinación de las creencias en enunciados probabilistas, sobre la base de otras creencias (y experiencias). Tanto las creencias como los objetivos, *creen* y g, entran en la determinación de nuestra conducta, pero entran de maneras diferentes, incrustadas en diferentes posiciones funcionales. (¿Se modificaría esta conclusión si incorporáramos explícitamente la determinación [parcial] de la *creencia* en términos de teoría de la decisión de la regla 2'?)

51. Ignorar estas posibilidades a la hora de revisar probabilidades ¿lleva a la violación de algunos de los axiomas de la teoría de la probabilidad, y por lo tanto, a la violación de las condiciones de coherencia? ¿U ocurre la revisión contextualista en $1 - \epsilon$, permaneciendo ϵ *terra incognita*? ¿O tiene el contextualista radical probabilidades que están también, como sus creencias, vinculadas a un contexto, no probabilidades invariantes a través de los contextos?

52. Amos Tversky y Daniel Kahneman, «Judgment under Uncertainty: Heuristics and Biases», reproducido en *Judgment under Uncertainty: Heuristics and Biases*, comp. Daniel Kahneman, Paul Slovic y Amos Tversky (Cambridge: Cambridge Univ. Press, 1982), págs. 3-20. Véase especialmente su discusión de la «heurística de la disponibilidad», págs. 11-14.

53. Para una discusión de la investigación sobre «perseverancia de la creencia después de haber sido desacreditada por la evidencia», véase Lee Ross y Craig Anderson, «Shortcomings in the Attribution Process: On the Origins and Maintenance of Erroneous Social Assessments», en *Judgment under Uncertainty*, comp. Kahneman, Slovic y Tversky, esp. págs. 148-152.

54. Los psicólogos se han preocupado por los efectos continuados de las falsedades que les cuentan a las personas con las que experimentan (para evitar la contaminación de los resultados del experimento), aun a pesar de que luego les dicen la verdad, pues las falsedades pueden tener efectos continuados, aun después de ser descubiertas. Nuestras presentes reflexiones plantean cuestiones acerca de la adecuación de experimentos en los que el psicólogo cuenta la *verdad* desde el comienzo (o de experimentos realizados en medios naturales, sin dar información previa sobre el propósito del experimento). El mero hecho de que alguien respondiera de esta forma *en un experimento,* y hablara sobre ello luego con algún investigador,

puede dar a esta particular información un relieve especial y, así, una fuerza no representativa en la modelación de las creencias posteriores de la persona acerca de su propio carácter y de sus propias capacidades. Si la verdad puede sesgar —no sólo las falsedades—, entonces los psicólogos podrían tener obligaciones adicionales para contrarrestar los efectos de sus intervenciones experimentales en las vidas de la gente.

55. Hay otros sesgos en la estimación de la evidencia que podrían reclamar corrección. Véase el conjunto de *Judgment under Uncertainty,* comp. Kahneman, Slovic y Tversky.

56. Se dispone de cierto material sugerente en la bibliografía psicológica. Véase ibíd., artículos 30-32. Y recuérdense los escritos de C. S. Peirce sobre la naturaleza autocorrectora de los procedimientos científicos.

57. El deseo de «distribución geográfica» en sus admisiones por parte de las universidades selectivas originó una discriminación de segundo nivel en la Universidad de Harvard en 1992. Su presidente, A. Lawrence Lowell, defendió abiertamente cuotas restringiendo el número de judíos que podían ser admitidos en el Harvard College —un caso de aplicación explícita de criterios diferentes para diferentes grupos—. Cuando esta declaración de manifiesta discriminación de primer nivel produjo un alboroto público, la Universidad de Harvard descubrió las virtudes de la «distribución geográfica». Está claro que este objetivo se añadió a los objetivos tradicionales para limitar las admisiones de solicitantes judíos, que tendían a concentrarse en grandes ciudades. El proceso de admisión de Harvard habría progresado de la discriminación de primer nivel a la discriminación de segundo nivel. Para un estudio detallado de esta historia, véase Penny Feldman, «Recruiting an Elite: Admission to Harvard College» (Tesis Doctoral, Harvard Univ., 1975). Véase también Alan Dershowitz y Laura Hart, «Affirmative Action and the Harvard College Diversity-Discretion Model: Paradigm of Pretext?», *Cardozo Law Review* 1 (1979): 379-424.

58. P. Bickel, Eugene Hummel y J.W. O'Connell, «Is There a Sex Bias in Graduate Admissions», *Science* 187 (1975): 398-404.

59. Otra inferencia demasiado apresurada de no discriminación (o de discriminación relativamente poco grave) basada en estadísticas la realiza Thomas Sowell, quien argumenta como sigue. Casi ningún blanco puede distinguir, o no se para a distinguir, entre distintos subgrupos de negros, de manera que puede esperarse que los discrimine a todos ellos por igual. Sin embargo, el ingreso medio de los negros procedentes de las islas del Caribe es igual al de los blancos americanos. Así, pues, ¿no son los rasgos culturales de los otros subgrupos negros, y no la discriminación por parte de los blancos, lo que hace que el ingreso promedio de estos otros subgrupos negros esté por debajo del promedio de los blancos? Véase Thomas Sowell, *Civil Rights: Rhetoric or Reality?* (Nueva York: William Morros, 1984), págs. 77-79.

Sin embargo, hay subgrupos blancos cuyo ingreso promedio es superior al promedio blanco general —por ejemplo, los descendientes de escandinavos y los judíos—. Quizá los negros procedentes de las islas tendrían también un ingreso superior al promedio blanco si no fuera por la discriminación. Quizás haya discriminación contra todos los negros, lo que mantendría su ingreso por debajo del que tendrían si no hubiera discriminación. El que haya un subgrupo negro *al nivel* del promedio blanco no permite inferir que no haya discriminación contra todos los negros.

60. La rectificación de esta arbitrariedad de segundo nivel —no tiene por qué ser discriminación— debería distinguirse de otro objetivo que a veces se propone: fortalecer la autoimagen o la imagen externa de determinadas minorías que en los Estados Unidos están (o se sienten) oprimidas por la vía de incluir sus obras en el canon que se enseña. Obsérvese que puede haber un *propósito* educativo en la inclusión de algunos escritores procedentes de estos grupos, aun en el caso de que

sus méritos artísticos sean menores que los de los más grandes escritores que podrían incluirse. Ello no obstante, podría tratarse de escritores más sensibles y dotados que la inmensa mayoría de los estudiantes que habrían de leerlos. Exponerlos a la lectura de estudiantes que aún no reconocen que algunas mujeres y miembros de grupos minoritarios son mucho más inteligentes, sensibles y dotados que ellos mismos serviría a una importante función educativa.

En el último capítulo de *Philosophical Explanations*, describo un proceso de etapas alternativas de valor y significado: establecer unidades, extenderlas hasta que conecten entre sí e incluyan una diversidad adicional de material que rompa esas unidades, establecer nuevas y más amplias unidades, extendiéndolas hasta que conecten, etc. Los partidarios de una unidad (amenazada) presente podrían entender con provecho el «multiculturalismo» como parte de este proceso en curso, no como su etapa final.

61. Véase Robert Nozick, *The Examined Life* (Nueva York: Simon and Schuster, 1989), págs. 76-83, y Oliver Williamson, «Calculativeness, Trust and Economic Organization» (marzo 1992, manuscrito de una ponencia presentada a la Conference on Law and Economics, Univ. of Chicago Law School, abril 1992). Que alguien se abstenga de calcular cuán exactamente digno de confianza es un amigo, y se limite a confiar en él, no implica que una contraevidencia suficientemente robusta no pudiera bastar para convencerle de que ese amigo no es digno de confianza.

NOTAS AL CAPÍTULO 4

1. ¿No podría haber, sin embargo, concatenaciones o extensiones de esas relaciones, concatenaciones y extensiones que constituirían razones aun cuando nuestras mentes no pudieran reconocerlas? ¿Podría la relación de razón ser recursivamente enumerable pero no recursiva?

2. Véase Nelson Goodman, *Fact, Fiction, and Forecast* (Cambridge, Mass.: Harvard Univ. Press, 1955), págs. 65-66, en donde se plantea una cuestión parecida para la concepción *a priori*.

3. *Philosophical Explanations* (Cambridge, Mass.: Harvard Univ. Press, 1981), págs. 248-253. Otros han presentado concepciones en las que el grado de apoyo es contingente. Cuánto apoyo signifique para la hipótesis de que todos los P son Q un número determinado de casos en los que los P son Q dependerá de la creencia existente sobre cuán variables tienden a ser los P en relación con los Q, esto es, sobre el espectro de variación en el tipo (relevante) de cosa que es P con respecto al tipo (relevante) de cosa que es Q. Véase John Holland, Keith Holyoak, Richard Nisbett y Paul Thagard, *Induction: Processes of Inference, Learning, and Discovery* (Cambridge, Mass.: M.I.T. Press, 1986), págs. 232-233, una concepción anticipada por Norman Campbell en *What Is Science?* (1921; reimp. Nueva York: Dover, 1952), págs. 63-64.

4. A la vista de los debates recientes sobre adaptacionismo, sería deseable que esta hipótesis no exigiera abiertamente demasiada concreción en la selección evolucionaria de rasgos del cerebro. Véase Stephen Jay Gould y Richard Lewontin, «The Spandrals of San Marcos and the Panglossian Paradigm: A Critique of the Adaptationist Programme», *Proceedings of the Royal Society of London*, B 205 (1979): 581-598; véanse también los varios ensayos que discuten la optimalidad en *The Latest on the Best: Essays on Evolution and Optimality*, comp. John Dupre (Cambridge, Mass.: M.I.T. Press, 1987), caps. 4-9.

5. Leda Cosmides y John Tooby, «Are Humans Good Intuitive Statisticians After All?» (en prensa); Leda Cosmides, «The Logic of Social Exchange: Has Natural Se-

lection Shaped How Humans Reason?», *Cognition* 31 (1989): 187-276; Leda Cosmides y John Tooby, «From Evolution to Behavior», en *The Latest on the Best,* comp. Dupre; John Tooby y Leda Cosmides, «The Psychological Foundations of Culture», en *The Adapted Mind,* comp. J. Bardow, L. Cosmides y J. Tooby (Nueva York: Oxford Univ. Press, en prensa), págs. 19-136.

6. Véase Daniel Dennett, *Consciousness Explained* (Boston: Little, Brown, 1991) [trad. cast.: *La conciencia explicada,* Barcelona, Paidós, 1995], págs. 184-187, para una discusión del efecto Baldwin.

7. W.V. Quine, «Truth by Convention» (1936), reproducido en W.V. Quine, *The Ways of Paradoxe* (Cambridge, Mass.: Harvard Univ. Press, 1976), págs. 77-106.

8. La distinción entre evidencia como una conexión fáctica y como una conexión evidente (casi) *a priori* es análoga a la distinción entre la racionalidad en tanto que surgida de un proceso fácticamente fiable y la racionalidad en tanto que constituida por un cierto tipo de mezcolanza densa de enunciados, argumentos e inferencias solapados y conectados. En ambos casos tenemos un aspecto fáctico distinguido de un aspecto racional; en ambos queremos que los dos vayan de consuno y nos sentimos incómodos con un aspecto racional no arraigado en una conexión fáctica. Cuando creemos que el aspecto racional está ligado al fáctico, nos sentimos cómodos afirmando que una ejemplificación racional es valiosa en sí misma. Mas cuando los dos aspectos se ven desligados, cuando un modo de racionalidad no parece ya reflejar los hechos o constituir una vía de obtenerlos —como ocurrió con la tradición de las disputas escolásticas—, entonces ese modo pierde su halo, ya no parece bello o intrínsecamente valioso.

Valga esta comparación: en ética es mucho más fácil sentirse cómodo con una posición deontológica cuando las consecuencias de la acción requerida o correcta parecen también razonablemente buenas.

9. Si el razonamiento inductivo *es* racional, entonces hay un argumento racional de este tipo, el argumento inductivo; esto se desecha por circular. Así, se supone que el problema consiste en dar apoyo a una parte de la Razón, el razonamiento inductivo, con *otras* partes de la razón, es decir, de una manera no circular.

10. Véase también la nota 49 del tercer capítulo.

11. Kant, *Critique of Pure Reason,* trad. Norman Kemp Smith (Londres: Macmillan, 1933) [trad. cast.: *Crítica de la razón pura,* Madrid, Alfaguara, 1994]. Prefacio a la segunda edición.

12. Compárese W.V. Quine, *Word and Object,* (Cambridge, Mass.: M.I.T. Press, 1960), cap. 2.

13. Kant, *Critique of Pure Reason* [trad. cast.: *Crítica de la razón pura,* Madrid, Alfaguara, 1994]. Prefacio a la segunda edición.

14. Véase Robert Nozick, «Experience, Theory and Language», en *The Philosophy of W.V. Quine,* comp. Lewis Hahn (La Salle, Ill.: Open Court, 1986), págs. 340-341, y Stephen Stich, *The Fragmentation of Reason* (Cambridge, Mass.: M.I.T. Press, 1990), págs. 60-63. Stich llega a argumentar que, a causa de esas diferencias en los costes del error, el mecanismo cognitivo seleccionado positivamente por la evolución quizá no sea el detector más fiable de la verdad cuando esta menor fiabilidad se vea rebasada por las restantes virtudes del mecanismo.

15. La teoría de la genética de poblaciones, plenamente desarrollada en el detalle, tiene la siguiente estructura, según nos enseña Richard Lewontin (Lewontin, *The Genetic Basis of Evolutionary Change* [Nueva York: Columbia Univ. Press, 1974] [trad. cast.: *La base genética de la evolución,* Barcelona, Omega, 1979], cap. 1, esp. págs. 12-15, de donde tomo la exposición del resto de este párrafo). Consiste en descripciones genotípicas *G1* y *G2* de la población en los tiempos *t1* y *t2*, y en leyes de trans-

formación que van de una a otra: un conjunto de leyes epigenéticas que dan la distribución de fenotipos resultantes del desarrollo de los varios genotipos en varios medios; leyes de apareamiento, de migración y de selección natural que transforman la disposición fenotípica de una población en el lapso de una generación; el conjunto de relaciones epigenéticas que permiten inferencias acerca de la distribución de genotipos que se corresponde con cualquier distribución de fenotipos; y las reglas genéticas (de Mendel y Morgan) que predicen la disposición de los fenotipos en la próxima generación, producida a partir de la gametogénesis y de la fertilización, dada una disposición de genotipos parentales. Los genotipos y los fenotipos son variables de estado; la teoría de la genética de poblaciones proyecta un conjunto de genotipos en un conjunto de fenotipos, transforma éstos en otros fenotipos y luego retroproyecta el resultado en los genotipos, que son entonces transformados para producir la disposición genotípica de la siguiente generación.

Trabajando en esta estructura, Elliott Sober construye la teoría evolucionaria como una teoría de fuerzas que actúan sobre un estado de equilibrio de fuerza cero definido por la ecuación de Hardy-Weinberg. (Esa ecuación dice que, después de la primera generación, la proporción de alelos en cada *locus* dentro de una población se mantendrá constante a menos que actúen sobre ella fuerzas externas, y da una fórmula para esa proporción.) La teoría evolucionaria determina el modo en que ese equilibrio cambia cuando está sometido a varias fuerzas (selección, mutación, migración, deriva genética) que actúan por separado y en combinación. (Véase Elliott Sober, *The Nature of Selection* [Cambridge, Mass.: M.I.T. Press, 1984], cap. 1.) Sin embargo, como observa John Beatty, la ley de Hardy-Weinberg es una consecuencia de la herencia mendeliana, y los mecanismos de la misma —reproducción sexual, cruces macho-hembra y hembra-macho con resultados equivalentes, mecanismos que satisfacen la ley de la segregación y del surtido independiente— son ellos mismos un producto de la evolución. Si la teoría evolucionaria ha de explicar también cómo surgió la herencia mendeliana, el enfoque que da Sober a la misma no puede ser completo. (Véase John Beatty, «What's Wrong with the Received View of Evolutionary Theory?», en *Proceedings of the P.S.A., 1980*, comp. Peter Asquith y Ronald Giere [East Lansing, Mich.: Philosophy of Science Association, 1980], vol. 2.)

Podemos generalizar el enfoque de Sober y evitar esa dificultad viendo la teoría evolucionaria como una teoría histórica que describe una secuencia de *diferentes* estados de equilibrio de fuerza cero. Cada uno de estos estados va emparejado con un mecanismo de herencia (¿y en la primera generación?), y la teoría de cada uno de esos estados determina las fuerzas que pueden romper ese equilibrio y las leyes de ese rompimiento. Algunos rompimientos generarán un nuevo mecanismo de herencia, el cual, una vez en acto, dará lugar a su propio estado de fuerza cero, a las leyes de rompimiento, y así sucesivamente. Obtenemos de este modo una narración histórica de la secuencia de estados de fuerza cero y de los mecanismos que van con ellos, dando lugar cada estado al siguiente de acuerdo con las leyes de transformación vinculadas con este estado y con su mecanismo. Con cada nuevo estado de equilibrio puede venir una nueva lista de fuerzas desviantes y nuevas leyes sobre el modo de operar de esas fuerzas. Así, pues, en el esquema de Lewontin, el resultado de una transformación puede ser un nuevo estado natural (de fuerza cero) con distintas fuerzas y leyes desviantes. Pero aun cuando el estado de equilibrio, el particular mecanismo de herencia y las fuerzas desviantes pueden cambiar todos —en este sentido, la teoría es radicalmente histórica—, lo que hace que esto sea una historia evolucionaria ininterrumpida es el papel constante desempeñado por las variaciones heredables de adaptación.

16. Susan Mills y John Beatty, «The Propensity Interpretation of Fitness», reproducido en *Conceptual Issues in Evolutionary Biology*, comp. Elliott Sober (Cambridge, Mass.: M.I.T. Press, 1984), págs. 36-57.

17. Véase John Beatty y Susan Finsen, «Rethinking the Propensity Interpretation», en *What the Philosophy of Biology Is: Essays for David Hull*, comp. Michael Ruse (Dordrecht: Kluwer, 1989), págs. 17-30.

18. Robert Brandon reconoce que un valor del número de descendientes en la próxima generación superior al esperado resulta relevante —una varianza incrementada puede resultar selectivamente desventajosa—. Por eso propone medir la adaptación substrayendo del número esperado de descendientes alguna función de la varianza. Véase Robert Brandon, *Adaptation and Environment* (Princenton: Princeton Univ. Press, 1990), págs. 39-77. ¿Pero *qué* particular función hay que substraer? Beatty y Finsen sostienen que no es simplemente una cuestión de media y varianza; el sesgo de una distribución es relevante también. Puesto que las estadísticas particulares son un componente de la estrategia del organismo en un medio, la adaptación en general no debería ser identificada con *una* de esas estadísticas. Véase Beatty y Finsen, «Rethinking the Propensity Interpretation», págs. 17-30.

19. Véase Ernest Nagel, *The Structure of Science* (Nueva York: Harcourt, Brace and World, 1961) [trad. cast.: *La estructura de la ciencia*, Barcelona, Paidós., 1991], págs. 401-428. Nagel siguió al biólogo G. Sommerhoff, *Analytical Biology* (Londres, 1950).

20. Larry Wright, «Functions», *Philosophical Review* 82 (1973), reproducido en *Conceptual Issues in Evolutionary Biology*, comp. Sober, págs. 347-368.

21. Christopher Boorse, «Wright on Functions», reproducido en *Conceptual Issues in Evolutionary Biology*, com. Sober, págs. 369-385.

22. Estos ejemplos proceden de Peter Godfrey-Smith, pero él ofrece una interpretación distinta de por qué no puede decirse que sean funciones.

23. Esto no significa necesariamente que en el pasado hubiéramos conseguido hallar y probar, por ejemplo, los principios de inducción, la existencia de otras mentes y del mundo externo, y que, al habernos ido especializando desde entonces en el trabajo en equipo con esos hechos, hayamos acabado perdiendo la capacidad para justificarlos o probarlos.

24. Rudolf Carnap, «Testability and Meaning», *Philosophy of Science* 3 (1936): 419-471 y 4 (1937): 1-45.

25. L.J. Savage, *The Foundations of Statistics* (Nueva York: John Wiley, 1954).

26. El argumento del libro holandés, que se discute más abajo, no proporciona una razón independiente de este tipo. A lo sumo, dice por qué, si hay probabilidades personales, de acuerdo con las cuales actúa siempre una persona, estas probabilidades deberían satisfacer los axiomas usuales de la teoría de la probabilidad; no dice por qué debe o debería haber tales probabilidades que siempre guían a las elecciones (de apuesta).

27. Aunque ya desarrollé todos los detalles de este párrafo en mi *The Normative Theory of Individual Choice* (1963; reimp. Nueva York: Garland Press, 1990), págs. 159-172 y 246-250, no fue sino después de hablar con Hilary Putnam y leer un ensayo reciente suyo, «Pragmatism and Moral Objectivity» (en prensa) —en el que insiste independientemente en la cuestión de por qué debe actuarse de acuerdo con lo más probable—, que me di cuenta de que se trataba de un problema serio, no meramente de una dificultad remilgadamente artificiosa. (Otro asunto relacionado —por qué debería actuarse basándose en la certeza antes que en la probabilidad—, *acaso* pueda resolverse con consideraciones de dominación, *si* es que no hay aquí petición de principio.)

28. Consideremos, por último, un problema acerca de la racionalidad misma. Sea el enunciado *R*: Cree el enunciado *p* (o haz la acción *A*) si y sólo si *p* (o *A*) es demostrablemente racional. (O podría haberse demostrado que lo es cuando se adquirió esta creencia. Para mantenerla, quizá baste con que no se demuestre que es irracional.) Tenemos robustos fundamentos inductivos —es decir, buenas razones— para suponer que no puede demostrarse que *R* mismo sea racional —nadie lo ha conseguido hasta ahora, a pesar de que se han hecho esfuerzos muy serios—. Supongamos que es así. Entonces, si *R* es verdadero, ustedes no deberían creerlo, pues no puede demostrarse que es racional. Entonces hay al menos una verdad a la que no les conduce la racionalidad. (¿Y si hay una, por qué no más?) Por otro lado, si *R* es falso, entonces hay algo que tiene que ser creído (o hecho) a pesar de que no puede demostrarse que sea racional, *o* hay algo que no puede ser creído (o hecho) a pesar de que puede demostrarse que es racional. En cualquiera de los casos, la racionalidad parece limitada. Hay robustos fundamentos inductivos para creer que, una vez definida alguna noción concreta de racionalidad, no puede demostrarse que *R* sea racional. De acuerdo con sus propios criterios, la racionalidad no ha conseguido (hasta ahora) justificarse a sí misma. De modo que parece racional para nosotros creer que no puede demostrarse que *R* sea racional y, por lo tanto, si *R* es verdadero, no creer *R*. (¿Cómo se vería afectada la situación si se substituyera *R* por las reglas 1 y 2'?)

29. Véase Donald Norman, *The Psychology of Everyday Things* (Nueva York: Basic Books, 1988), cap. 3.

30. Hubert Dreyfus ha sostenido que el proyecto de la inteligencia artificial se ha encontrado con dificultades debido a la naturaleza incrustada y corpórea de nuestra racionalidad. Véase su *What Computers Can't Do: A Critique of Artificial Reason* (Nueva York: Harper and Row, 1972).

31. En esta cuestión ha insistido el trabajo de Nelson Goodman sobre «grue» y «bleen», insólitos predicados que casan con la conducta pasada de las cosas *green* (verdes) y *blue* (azules), pero que divergen de los predicados usuales en el futuro. Véase Goodman, «A Query on Confirmation», *Journal of Philosophy* 43 (1946): 383-385, y *Fact, Fiction, and Forecast*, págs. 73-83. Lo mismo puede establecerse haciendo pasar a dos curvas diferentes a través de los mismos datos de puntos del pasado. ¿Cuál de esas regularidades continuará manteniéndose?

32. Resulta instructivo comparar las observaciones de Wittgenstein en *On Certainty* (Oxford: Basil Blackwell, 1969) [trad. cast.: *Sobre la certeza*, Barcelona, Gedisa, 1987], par. 83, 88, 94, 103, 105, 152, sobre «el marco de referencia», las «proposiciones que están firmes para mí» y lo que «está anclado en todas mis cuestiones y respuestas, tan anclado que no puedo tocarlo», con nuestra hipótesis de una evolución que nos inculca, en calidad de herencia filogenética, hechos estables de medios pasados. Ello no obstante, Wittgenstein no considera los modos en que los componentes inculcados de este marco podrían operar para inducirse recíprocamente cambios.

33. Véase Douglass North, *Institutions, Institutional Change and Economic Performance* (Cambridge: Cambridge Univ. Press, 1990); Andrew Schotter, *The Economic Theory of Social Institutions* (Cambridge: Cambridge Univ. Press, 1981); Oliver Williamson, *The Economic Instituions of Capitalism* (Nueva York: Free Press, 1985); Margaret Levi, «A Logic of Institutional Change», en *The Limits of rationality*, comp. Karen Schweers Cook y Margaret Levi (Chicago: Chicago Univ. Press, 1990), págs. 383-401; Thrainn Eggertsson, *Economic Behavior and Institutions* (Cambridge: Cambridge Univ. Press, 1990); Harold Demsetz y Armen Alchian, «Production, Information Costs and Economic Organization», *American Economic Review* 62 (1972):

777-795; Michael Jensen y William Meckling, «Theory of the Firm: Managerial Behavior, Agency Costs and Ownership Structure», *Journal of Financial Economics* 3 (1976): 305-360 (reimp. en *Economic and Social Institutions*, comp. Karl Brunner [Boston: Martinus Nijhoff, 1979], págs. 163-231); E. Furbotn y S. Pejovich, «Property Rights and the Behavior of the Firm in a Socialist State», en *The Economics of Property Rights*, comp. Furbotn y Pejovich (Cambridge: Ballinger, 1974), págs. 227-251; Dennis Mueller, *Public Choice II* (Cambridge: Cambridge Univ. Press, 1989); Gary Becker, *The Economic Approach to Human Behavior* (Chicago: Univ. of Chicago Press, 1976); James Coleman, *Foundations of Social Theory* (Cambridge, Mass.: Harvard Univ. Press, 1990); Richard Swedberg, *Economics and Sociology* (Princeton: Princeton Univ. Press, 1990).

34. Al revés, ¿podría una institución funcionar como si maximizara alguna función-objetivo por la concatenación de la conducta de los individuos que alberga y sin que éstos trataran de maximizar ésta o alguna otra función-objetivo?

35. Véase Richard Dawkins, *The Selfish Gene* (Oxford: Oxford Univ. Press, 1976) [trad. cast.: *El gen egoísta*, Cerdanyola, Labor, 1979].

36. Gary Becker avanza esta tesis en *A Treatise on the Family* (Cambridge, Mass.: Harvard Univ. Press, 1981) [trad. cast.: *Tratado sobre la familia*, Madrid, Alianza, 1987], pág. 102, y cita la literatura en que se basa.

37. «Esta adaptación [de las acciones del hombre] a las circunstancias generales que le rodean tiene su origen en la observación de reglas que no ha formulado él mismo y que a menudo no conoce explícitamente... nuestras acciones [están] gobernadas por reglas adaptadas al tipo de mundo en el que vivimos, esto es, a circunstancias de las que no somos conscientes y que, sin embargo, determinan la pauta de nuestras acciones exitosas.» Las reglas mismas «han evolucionado por un proceso de selección en la sociedad en la que vive, y son así, pues, el producto de la experiencia de generaciones.» F.A. Hayek, *Law, Legislation and Liberty*, vol. 1: *Rules and Order* (Chicago: Univ. of Chicago Press, 1973); págs. 11-12. El proceso descrito por Hayek es un proceso de selección de grupo. «Esas reglas de conducta no se han desarrollado, pues, como las condiciones reconocidas para el logro de un propósito conocido, sino que han evolucionado porque los grupos que las secundaban tuvieron más éxito y desplazaron a otros grupos» (pág. 18).

38. E.O. Wilson, *Sociobiology* (Cambridge, Mass.: Harvard Univ. Press, 1975) [trad. cast.: *Sociobiología*, Barcelona, Omega, 1980], pág. 145. El capítulo 7 del libro de Wilson constituye una elaboración de este tema, distinguiendo entre distintos niveles de respuesta a lo largo de distintos períodos de tiempo: organísmico, ecológico y evolucionario.

Para cada nivel de frecuencia de cambio, podría haber mecanismos adaptados para responder a cambios de (aproximadamente) esta frecuencia, produciendo cosas o entidades que perduraran aproximadamente a lo largo de este lapso, con modificaciones inducidas por reglas de retroalimentación apropiadas. Podemos distinguir estos casos: la cosa producida casa con el tipo de constante con la que está engranada; la constante ha cambiado y la cosa misma producida es cambiante con objeto de casar con la nueva constante —está en vías de alcanzar un nuevo equilibrio—; la constante está cambiando a una tasa de cambio tan elevada que rebasa la capacidad del mecanismo de retroalimentación para adaptarse a ella —la nueva cosa producida no casará con la constante que se da ahora—.

39. Un argumento, según el cual el proceso selectivo *es* severo y adecuado, ¿no estaría sesgado por la imagen que de sí misma ofrece la sociedad a sus miembros? (Compárese esto con la noción de ideología de los marxistas.) ¿Existe ese sesgo por-

que refleja alguna verdad robusta y duradera sobre la vida social, o acaso porque sirve al mantenimiento de este particular tipo de sociedad y contribuye a sostener la dominación de algún grupo particular dominante de la misma?

40. Véase Richard Dawkins, *The Blind Watchmaker* (Nueva York: W.W. Norton, 1986) [trad. cast.: *El relojero ciego*, Cerdanyola, Labor, 1989], págs. 77-86.

41. Además, según señala Paul David, la disposición de las teclas en el teclado de las máquinas de escribir corrientes en el mundo de habla inglesa es tecnológicamente ineficiente, pero dada la inversión realizada ya en equipo y en pericia mecanográfica, un cambio de teclado resultaría económicamente ineficiente. Por lo tanto, aun cuando un método de escalar cumbres puede llevar a un óptimo global, este óptimo puede tener también defectos, el remedio de los cuales resultaría ineficiente a causa de los ajustes históricos a que se procedió para conseguir su encaje. Véase Paul David, «Clio and the Economics of QWERTY», *American Economic Review* 75 (1985): 332-337.

42. Algunos han sostenido que el tipo de relaciones industriales habituales en Japón debería ser un modelo para los Estados Unidos, un argumento que se considera muy seriamente a causa de los éxitos japoneses en la competición económica internacional. Análogamente, la disposición observada en la Europa del Este y en la antigua Unión Soviética a realizar experimentos que implican grandes cambios hacia un capitalismo de mercado procede de la demostrada y visible superioridad del mundo capitalista en punto a prosperidad.

NOTAS AL CAPÍTULO 5

1. Obsérvese que ya tenemos *dentro* de la red dos piezas del dispositivo; la fórmula estándar de Bayes, que usa probabilidades condicionales evidenciales; y la versión causalizada, que usa probabilidades causales. ¿Hay algún modo de combinarlas, añadiendo además consideraciones simbólicas? A modo de aproximación tentativa —la red procesadora real tendrá una descripción mucho más complicada—, podríamos concebir el grado de credibilidad de *h* sobre la base de *e*, cred(*h*,*e*), como una suma ponderada de la *ratio* bayesiana causalizada (con condicionales subjuntivos), la *ratio* bayesiana estándar (con probabilidades condicionales) y un componente simbólico, sim(*h*,*e*). (No está claro, empero, cuál habría de ser el componente simbólico adecuado —qué simboliza creer *h* sobre la base de *e*—. ¿El grado en que simboliza creer la verdad?) ¿Y podrían esos pesos ser los mismos pesos que usamos también en nuestra teoría de la decisión, en nuestra fórmula *VD*? Una persona usará entonces *estos* valores de credibilidad para eliminar, como indignos de ser creídos, algunos enunciados; aquellos que compitan con algún enunciado incompatible que tenga un valor de credibilidad mayor. Y entonces la persona usará las reglas siguientes para decidir cuál de esos enunciados admisibles ha de creer.

2. David Hume, *A Treatise of Human Nature*, ed. L.A. Selby-Bigge (1888; Oxford: Oxford Univ. Press, 1958) [trad. cast.: *Tratado de la naturaleza humana*, Madrid, Editora Nacional, 1981], lib. II, pt. III, sec. III, pág. 416. Hume prosigue: «No es contrario a la razón para mí elegir mi ruina total con objeto de evitar la menor molestia a un hindú o persona completamente desconocida. Ni siquiera es más contrario a la razón preferir a sabiendas para mí un bien menor a un bien mayor».

3. Véase John Von Neumann y Oscar Morgenstern, *Theory of Games and Economic Behavior*, 3.ª ed. (Princeton: Princeton Univ. Press, 1953), apéndice; R.D. Luce y Howard Raiffa, *Games and Decisions* (Nueva York: John Wiley, 1957), págs. 12-38.

4. Alternativamente, se podría tratar de describir un *proceso* (normativo) de for-

mación y modificación de creencias, aislada o conjuntamente tomadas, y ver si ese proceso, llevado a cabo indefinidamente, resultaría en una función de utilidad de Von Neumann-Morgenstern. La teoría de la utilidad de Von Neumann-Morgenstern podría entonces entenderse como una descripción de estado-final del resultado de un proceso particular, al menos en el límite. Podría ocurrir entonces que no siempre tuviéramos una razón particular para llevar este proceso al límite.

5. ¿Deberíamos distinguir dos concepciones de los deseos exigidos por la racionalidad; deseos que resulta racional tener y deseos que resulta racional tener *si* la racionalidad misma es deseable o valiosa? ¿Podría argüirse que lo segundo no puede descartarse sin tornar confusa la cuestión de por qué es deseable tener los deseos que resulta racional tener?

6. Para una discusión de las preferencias de segundo orden, véase Harry Frankfurt, «Freedom of the Will and the Concept of the Person», *Journal of Philosophy* 68 (1971): 5-20; Amartya Sen, «Choice, Orderings and Morality», reproducido en su *Choice, Welfare and Measurement* (Oxford: Basil Blackwell, 1982), págs. 74-83; y Richard Jeffrey, «Preferences among Preferences», *Journal of Philosophy* 71 (1974): 377-391. Véase también Gilbert Harman, «Desired Desires», en *Value, Welfare, and Morality*, comp. R. Frey y C. Morris (en prensa).

7. William Talbott y Amartya Sen me sugirieron independientemente este punto.

8. Para una discusión de las dificultades que entraña perfilar esta disposición, véase Robert Nozick, *The Normative Theory of Individual Choice* (1963; reimp. Nueva York: Garland Press, 1990), págs. 39-48, 70-78.

9. Formular esta condición como una presunción que vale en ausencia de razones para que no valga evitar las objeciones que planteé en mi «On the Randian Argument», *The Personalist* 52, n. 2 (1971): 285-286, contra un principio más fuerte.

10. Debo esta idea sobre la identidad a Howard Sobel.

11. Una de las señales de que un deseo o una creencia es irracional es que no consigue pasar los controles y la modificación holísticos. Sólo se tiene en pie considerado en sí mismo, es resistente a la integración con otros deseos o creencias. Véase David Shapiro, *Neurotic Styles* (Nueva York: Basic Books, 1965). Conjeturo que lo mismo ocurriría en el caso de la sugestión posthipnótica; no se funde con ni llega a modificarse en el seno de la red holista de creencias y deseos.

12. Véase mi *Philosophical Explanations* (Cambridge, Mass.: Harvard Univ. Press, 1981), págs. 348-352, 714-716. Algunos autores han formulado condiciones adicionales que relacionan preferencias y deseos con conocimiento; no sólo con el conocimiento de sus causas, sino de sus consecuencias y de sus interrelaciones con cualquier otra cosa. Véase Richard Brandt, *A Theory of the Good and the Right* (Oxford: Oxford Univ. Press, 1979) [trad. cast.: *Teoría ética*, Madrid, Alianza, [2]1994], págs. 110-129, 149-162; para una crítica, véase Allan Gibbard, *Wise Choices, Apt Feelings* (Oxford: Oxford Univ. Press, 1990), págs. 18-22.

13. Para una discusión extremadamente iluminadora de varios asuntos suscitados en torno a los objetivos y sus funciones, véase Michael Bratman, *Intention, Plans, and Practical Reason* (Cambridge, Mass.: Harvard Univ. Press, 1987). Bratman discute muchos de estos asuntos bajo el tópico de «intenciones».

14. Véase Helmut Jungermann, Ingrid von Ulardt y Lutz Hausmann, «The Role of the Goal for Generating Actions», en *Analysing and Aiding Decision Processes*, comp. P. Humphreys, O. Svenson y A. Vari (Amsterdam: North Holland, 1983), esp. págs. 223-228.

15. Véase mi *The Examined Life* (Nueva York: Simon and Schuster, 1989), págs. 40-42.

16. ¿Hay algo ulterior que está en una relación con los objetivos como la que éstos guardan con los deseos y los deseos, con las preferencias, entrañando aún otro nivel de procesamiento y filtraje?

17. Recuérdese el tratamiento que da Isaac Levi a la creencia como un criterio de posibilidad seria, de manera que no resulta necesario asignar probabilidades o considerar situaciones en las que la creencia es falsa (véanse págs. 136-137, más arriba). La regla de Levi para llegar a creer algo puede decidir por una pequeña diferencia marginal; pero una vez algo se ha convertido en una *creencia*, los efectos son de largo alcance. Mientras que, en cambio, si regresamos a la situación anterior en la que esa regla generó la creencia, su diferencia con otra hipótesis, con otra posible creencia, habría sido muy pequeña, aparentemente no suficiente para arrojar directamente una tan gran diferencia de efectos.

18. Véase Henry Montgomery, «Decision Rules and the Search for a Dominance Structure», en *Analysing and Aiding Decision Processes*, comp. Humphreys, Svenson y Vari, págs. 343-369, y «From Cognition to Action», en *Process and Structure in Human Decision Making*, comp. Henry Montgomery y Ola Svenson (Nueva York: John Wiley, 1989), págs. 23-49. Montgomery considera que la regla de la maximización de la utilidad esperada es harina de otro costal, pues toma en cuenta toda la información. Pero obsérvese que la fórmula es un modo de combinar (¿de amalgamar?) la información en un atributo, la Utilidad Esperada, lo que le permite decir que una acción bate y domina a otra en todos los atributos relevantes, pues ahora sólo hay *un* atributo tal, la Utilidad Esperada, y (en ese nivel) ya no hay razones en contra de la acción maximizadora, ni razones en favor de otra acción.

19. Véase Bratman, *Intentions, Plans, and Practical Reason.*

20. Sería demasiado fuerte invertir la dirección de la condición; no toda preferencia que tengamos tiene por qué implicar el ser un tipo diferente de persona cuando la preferencia es satisfecha.

21. Obsérvese que esta condición no impide deseos que llevan necesariamente a creencias falsas o inconsistentes. (En el capítulo 3, en la sección «Reglas de racionalidad», dijimos que un procedimiento que lleva a un conjunto de creencias inconsistentes no es necesariamente irracional.) Harina de otro costal es si hay un deseo *de* tales creencias.

22. Estoy agradecido a Gilbert Harman por haberme llamado la atención sobre este punto.

23. Véase John Broome, *Weighing Goods* (Oxford: Basil Blackwell, 1991), págs. 100-107; Susan Hurley, *Natural Reasons* (Oxford: Oxford Univ. Press, 1989), caps. 4-6.

24. Obsérvese el paralelo entre esta discusión de la testabilidad de la teoría de la decisión, que interpretamos de modo que contenga un cuantificador existencial («existe un conjunto de aspectos que definen alternativas, tales que ...»), y nuestra anterior noción de adaptación, que empleaba una cuantificación existencial que cubría los rasgos heredables genotípicos.

25. Véase W.V. Quine, *Word and Object* (Cambridge, Mass.: M.I.T. Press, 1960), págs. 57-61; Donald Davidson, *Inquiries into Truth and Interpretation* (Oxford: Oxford Univ. Press, 1984), ensayos 9-13; David Lewis, «Radical Interpretation», en sus *Philosophical Papers*, vol. 1 (Oxford: Oxford Univ. Press, 1983), págs. 108-121; Ronald Dworkin, *Law's Empire* (Cambridge, Mass.: Harvard Univ. Press, 1986) [trad. cast.: *El imperio de la justicia*, Barcelona, Gedisa, 1988], cap. 2; Hurley, *Natural Reasons*, cap. 5.

26. Para distintas versiones de esta propuesta, véase David Lewis, «Radical Interpretation», págs. 108-118; Richard Grandy, «Reference, Meaning and Belief», *Journal of Philosophy* 70 (1973): 439-452.

27. Véase *Judgment under Uncertainty: Heuristics and Biases*, comp. Daniel Kah-

neman, Paul Slovic y Amos Tversky (Cambridge: Cambridge Univ. Press, 1982); Lee Ross y Richard Nisbett, *Human Inference* (Englewood Cliffs, N.J.: Prentice Hall, 1980); y Paul Thagard y Richard Nisbett, «Rationality and Charity», *Philosophy of Science* 50 (1983): 250-267, que discute las implicaciones de estos resultados psicológicos para la formulación de un principio de interpretación. Pero véase, a modo de contraste, la interpretación de la investigación de Tversky y Kahneman en Gerd Gigerenzer, «How to Make Cognitive Illusions Disappear», *European Review of Social Psychology* 2 (1991): 83-115, y en Leda Cosmides y John Tooby, «Are Humans Good Intuitive Statisticians After All?» (en prensa).

28. Thagard y Nisbett, «Rationality and Charity», menciona a los maestros del zen y a Hegel como ejemplos.

29. Véase Jack Goody y Ian Watt, «The Consequences of Literacy», *Comparative Studies in History and Society* 5 (1963): 304-305; Jack Goody, *The Domestication of the Savage Mind* (Cambridge: Cambridge Univ. Press, 1977) [trad. cast.: *La domesticación del pensamiento salvaje*, Torrejón de Ardoz, Akal, 1985], págs. 36-51, 74-111; Goody, *The Logic of Writing and the Organization of Society* (Cambridge: Cambridge Univ. Press, 1986) [trad. cast.: *La lógica de la escritura y la organización de la sociedad*, Madrid, Alianza, 1990], págs. 1-20, 171-185.

30. Véase Donald Davidson, «On the Very Idea of a Conceptual Scheme», en su *Inquiries into Truth and Interpretation* (Oxford: Oxford Univ. Press, 1984) [trad. cast.: *De la verdad y de la interpretación*, Barcelona, Gedisa, 1989], págs. 183-198.

31. Véase Susan Hurley, «Intelligibility, Imperialism, and Conceptual Scheme», *Midwest Studies in Philosophy* (en prensa).

32. Véase Alasdair MacIntyre, *Whose Justice? Which Rationality?* (Notre Dame, Ind.: Univ. of Notre Dame Press, 1988) [trad. cast.: *Justicia y racionalidad*, Barcelona, Ediciones Internacionales Universitarias, 1994].

33. Véase Dworkin, *Law's Empire*, págs. 46-68, 76-86.

34. Véase Stephen Jay Gould y Richard Lewontin, «The Spandrals of San Marcos and the Panglosian Paradigm: A Critique of the Adaptationist Programme», *Proceedings of the Royal Society of London*, B 205 (1979): 581-598. Daniel Dennett no sólo ha trazado la analogía entre la tarea interpretativa y la tarea de la explicación evolucionaria, sino que ha mantenido que ambas tareas son realmente idénticas y se dejan guiar por un supuesto de optimización. Véase Dennett, *The Intentional Stance* (Cambridge, Mass.: M.I.T. Press, 1987) [trad. cast.: *La actitud intencional*, Barcelona, Gedisa, 1991], págs. 237-321.

35. Para un temprano ejemplo de esa noción de utilidad condicional, véase mi *Normative Theory of Individual Choice*, págs. 144-154.

36. Sobre los puntos abordados en este párrafo, véase mi *Normative Theory of Individual Choice*, págs. 94-98.

37. De un modo independiente, Gilbert Harman ha objetado que las probabilidades pasadas no deberían ser vinculantes, pero él no considera el argumento intertemporal de libro holandés. Véase su «Realism, Antirealism and Reasons for Belief» (en prensa).

38. El bayesiano estricto puede esperar que la persona muestre también «buen juicio» en sus particulares probabilidades personales, pero no hay ninguna condición específica adicional para fijar eso.

39. Véase Paul Teller, «Conditionalization and Observation», *Synthese* 26 (1973): 218-258, que formula un argumento atribuido a David Lewis. Bas van Frassen observa que este argumento depende de que la persona viole la condicionalización siguiendo alguna otra regla determinada; en consecuencia, sostiene que la violación

de la condicionalización sólo es permisible mientras no siga regla alguna. Bas van Frassen, *Laws and Symmetry* (Oxford: Oxford Univ. Press, 1989), págs. 160-176.

40. Véase John Earman, *Bayes or Bust* (Cambridge, Mass.: M.I.T. Press, 1992), págs. 195-198. Más críticas a la condicionalización bayesiana pueden encontrarse en F. Bacchus, H.E. Kyburg, Jr. y M. Thalos, «Against Conditionalization», *Synthese* 85 (1990): 475-506.

41. Ésta es, según creo, la concepción de Nicholas Rescher: «La racionalidad consiste en la búsqueda inteligente de los objetivos adecuados». Rescher, *Rationality* (Oxford: Clarendon Press, 1988) [trad. cast.: *La racionalidad*, Madrid, Tecnos, 1993], pág. VII.

42. Herbert Simon y Allen Newell, *Human Problem Solving* (Englewood Cliffs, N.J.: Prentice Hall, 1972), págs. 71-105.

43. El más destacado exponente del modelo de problemas para la historia intelectual es Karl Popper. En su último escrito, Popper situó su enfoque en su contexto de los «tres mundos»: el mundo I de los objetos físicos, el mundo II de los estados de consciencia y el mundo III de los contenidos objetivos del pensamiento, mundo este último que incluye pensamientos científicos y poéticos (y, añade Popper, obras de arte). Entre los moradores del mundo III, dice Popper, están las situaciones de problemas, consistentes en un problema, su trasfondo (esto es, el lenguaje usado y las teorías incorporadas a la estructura del problema) y el marco de conceptos y teorías de que disponemos para enfrentarnos al problema. Además, el mundo III contiene argumentos críticos, sistemas teóricos y estado de discusión de un argumento. La historia de la ciencia, afirma Popper, debería ser no una historia de meras teorías, sino de situaciones de problemas y del modo en que éstas van modificándose a través de los intentos de solucionar los problemas —esos intentos son las teorías—. (Véase Karl Popper, «On the Theory of the Objective Mind», en su *Objective Knowledge* [Oxford: Oxford Univ. Press, 1972] [trad. cast.: *Conocimiento objetivo*, Madrid, Tecnos, ³1988], págs. 153-190, esp. pág. 177.) La comprensión histórica, dice Popper, se basa en el análisis de las relaciones del mundo III, no de los procesos de pensamiento del mundo II (pág. 178). El historiador del pensamiento estudiará productos intelectuales, sus rasgos estructurales, compatibilidades y relaciones teóricas, y estudiará asimismo esos productos como respuestas a la situación de problemas. Un «análisis situacional» es una reconstrucción idealizada de la situación de problemas en la que se halló el agente mismo, una reconstrucción que hace racionalmente comprensible (en la medida en que pueda hacerse) la acción o la teoría del agente, mostrando por qué resultaban éstas adecuadas a la situación desde el punto de vista del agente.

Esta adecuación a la situación ¿se entiende según viera el agente la *situación*, o según viera el agente la *adecuación*? (¿O según ambas?) La primera alternativa obligaría al historiador a describir la situación según la veía la persona, para luego tratar de mostrar que la acción o la teoría de la persona resultaba adecuada a esa situación, entendiera o no la persona misma la adecuación de este modo. Al proceder así, el historiador puede importar otros criterios de adecuación, los propios de su tiempo o cualesquiera criterios que considere correctos. Una cuestión interesante es la de saber si la respuesta del pensador en esta situación fue una respuesta correcta. Pero lo que esta persona trataba de hacer era llegar a una solución (en esta situación de problemas, tal como ella la veía) que fuera adecuada según *sus* criterios de adecuación, o según los criterios de su disciplina en aquel momento (según ella los entendiera), no según criterios posteriores, o según nuestros criterios. Para explicar por qué Galileo no aceptó las leyes de Kepler sobre el movimiento de los planetas, Popper dice que Galileo estaba justificado en no aceptar entonces las le-

yes de Kepler y en trabajar, en cambio, con una audaz supersimplificación (pág. 173). Ahora bien; esto casa con la metodología *popperiana;* pero sólo si casa con la de Galileo (y si Galileo la estaba siguiendo entonces), tendremos una explicación de por qué Galileo no aceptó esas leyes.

John Passmore describe un modo de hacer historia de la filosofía —el modo que a él le parece correcto— como historia *problemática.* Trata ese modo de entender los problemas a los que estaba haciendo frente un filósofo, las cuestiones a las que pretendía dar respuesta, y trata entonces de seguir los pasos de su construcción teórica como un intento de solventar esos problemas y de dar respuesta a esas cuestiones. (John Passmore, «The Idea of a History of Philosophy», *History and Theory* 4 [1964-1965]: 3-32). ¿Hasta qué punto son constantes los problemas y las cuestiones a los que se han enfrentado los filósofos a lo largo del tiempo? ¿Pueden ser las cuestiones lo bastante similares para que las respuestas propuestas para una de ellas cuenten también como posibles respuestas a otra cuando hay diferentes razones para plantearlas, aun siendo similares esas razones? Para que las *cuestiones* sean las mismas, ¿tiene también que ser la misma gama (implícita) de *respuestas* posibles, o solaparse en gran medida al menos? La cuestión «¿por qué esto?» a menudo no es sino una versión implícita de la cuestión «¿por qué esto y no esto otro?» Cuando dos períodos históricos se preguntan acerca del mismo «esto», pero contrastándolo con diferentes «estos otros», ¿puede decirse que sus cuestiones son suficientemente similares para que las respuestas a las mismas puedan competir o iluminarse mutuamente? La cuestión «¿cómo es esto posible?» es una versión tácita de la cuestión «¿cómo es esto posible dado que esto otro es verdad?». Cuando dos períodos históricos andan intrigados respecto de las posibilidades del mismo «esto», pero en relación a diferentes «estos otros» que parecen excluirlo, ¿están planteando la misma cuestión, y aun están hablando del mismo «esto»? Cuando dos teóricos se preguntan por la posibilidad del libre arbitrio, uno, acaso, porque da sentada la omnisciencia divina, y el otro, acaso, porque da por sentado el determinismo causal universal, ¿están investigando la misma cuestión o hablando de la misma cosa? Aun si los problemas no son constantes, las historias orientadas a problemas pueden estudiar el modo en que los teóricos del pasado estaban tratando de resolver *sus* problemas y por qué los problemas de los filósofos han ido cambiando a lo largo del tiempo.

En historia del arte, Michael Baxandall ha propuesto que entendamos al pintor como a alguien que se está enfrentando a un problema, de manera que su producto se interprete como la solución cumplida y concreta de ese problema. Para entender el producto, necesitamos reconstruir el problema específico, para cuya solución estaba diseñado el producto, y las circunstancias concretas que rodearon al hecho de que el pintor se planteara el problema. (Véase Baxandall, *Patterns of Intention* [New Haven: Yale Univ. Press, 1985] [trad. cast.: *Modelos de intención*, Madrid, Hermann Blume Central de Distribuciones, 1989].) Anteriormente, E.H. Gombrich había descrito la historia de la pintura figurativa en occidente como una serie de experimentos diseñados para resolver determinados y cambiantes problemas de acuerdo con una pauta de esquema y corrección. (Véase Gombrich, *Art and Illusion* [Nueva York: Pantheon, 1960] [trad. cast.: *Arte e ilusión*, Barcelona, Gustavo Gili, ^{2}1982]. Gombrich reconoce la influencia del pensamiento de Karl Popper.)

En un conocido manifiesto, el historiador inglés del pensamiento político Quentin Skinner propone un programa de investigación histórica que rechaza el modelo de problemas. No hay que entender, dice, a los teóricos políticos como si ofrecieran respuestas a cuestiones o posiciones perennes sobre tópicos intemporales, ni siquiera como si trataran de resolver problemas *intelectuales* del momento. Ocurre más bien que sus escritos son intervenciones en controversias concretas, y deberíamos entender

su propósito principal, su acto ilocucionario, como *haciendo eso*, es decir, prestando su apoyo a una de las partes de una determinada controversia social y política, argumentando en favor de la posición de esa facción, etc. (Véase Skinner, «Meaning and Understanding in the History of Ideas», *History and Theory* 8 [1969]: 3-53.) La intención del escritor es una intención determinada, específica, en una situación determinada. (Skinner concede que también pueden estudiarse otras cosas, pero considera central para su modo de hacer historia del pensamiento la identificación y el estudio de intervenciones determinadas en controversias concretas.)

Miles de personas han tomado, empero, diferentes partidos en cada controversia determinada. La razón de que estemos interesados en *esos* escritos no es que hayan tomado un partido, sino que han dicho algo interesante, algo que en verdad transciende esta o aquella controversia particular y tiene una validez más general. Si no fuera así, no sería una tarea tan delicada la de identificar la particular controversia en la que (supuestamente) el escritor pretendió intervenir. En realidad, cuál sea la controversia en la que se considere que el escritor ha tomado parte puede depender de la precisa datación de su escrito. Un año diferente, una controversia diferente, una intervención diferente.

Ni que decir tiene que casi siempre está en curso una que otra controversia social o política, de manera que no resulta sorprendente que los historiadores del pensamiento puedan hallar siempre una controversia en la que contextualizar un escrito. Si el escritor dice algo de alcance general, lo que diga tendrá implicaciones para varias controversias distintas posibles. El que diga algo en un momento determinado, con implicaciones para una controversia del momento, *no* significa que su intención (o su acto ilocucionario) sea tomar parte en *esta* controversia, y desde luego no que su intención sea *sólo* la de tomar un partido. Pues el escritor puede tratar de proponer una teoría o verdad general de relevancia y aplicación amplias. Su acto ilocucionario, si hay que introducir esta categoría, puede ser el de *teorizar*. Es posible que los teóricos de la política traten de decir cosas intemporales que se aplican a (muchos) otros contextos y tiempos, de manera que entenderlos como si hablaran sólo de un determinado contexto y de una determinada controversia sería distorsionar su propósito.

Aun en el caso de que coincidamos con el sociólogo o con el historiador en que uno de los propósitos de un autor era promover una determinada causa en una controversia, aún deberíamos preguntarnos por qué lo hizo con argumentos teóricos abstractos, ofreciendo principios generales. Para ganar a otros para su causa o para hacer más firme su posición no puede limitarse a declarar sus preferencias por este partido; tiene que producir razones que resulten convincentes para los demás. Esas razones podrían ser concretas, pero también pueden ser reflexiones teóricas generales, aplicables a un amplio espectro de casos, uno de los cuales vendría en apoyo del partido por él tomado. Si los otros casos en que resultan aplicables sus reflexiones generales son casos que la otra persona acepta ya, entonces (por razonamiento general) estos otros casos quedarán reclutados como evidencia en favor del juicio propuesto en el caso en cuestión.

Así, pues, aun si un autor pretende intervenir en una controversia particular, aun si su intención principal no es *teorizar*, *nosotros* estaremos interesados en su trabajo no porque intervenga en favor de una parte, sino porque consigue presentar una teoría general y posiblemente convincente que se aplica a un amplio espectro de casos, de situaciones históricas, etc. El alcance de *nuestro* interés tendrá que ver con el alcance de su éxito en punto a presentar una teoría general atractiva y aparentemente convincente de dilatada aplicación. (Recuérdese que hay miles de personas que simplemente se alinearon con una u otra de las partes de una disputa,

personas a las que no estudiamos con el mismo detalle.) Lo que a nosotros nos interesa del teórico, lo que nos lo hace importante, no es el hecho de que se alineara —*si* lo hizo—, sino la teoría que desarrolló. Aun si el autor no estaba tratando simplemente de teorizar, sí estaba tratando de buscar justificaciones sirviéndose de razonamientos abstractos y generales, de manera que no podemos entender lo que hacía el autor a menos que nos centremos en lo que *él* se centraba, a saber: en la estructura de razones que habían de venir en apoyo de una posición general, en la medida en que esa estructura afectaba a la adecuación y a la aceptabilidad de la posición tomada. Si el acto ilocucionario del escritor es de justificación, una de nuestras preocupaciones primordiales será la de investigar si, y en qué medida, *justificó*. La historia del pensamiento, pues, debe consistir en gran parte en historia de ideas, de teorías y de posiciones razonadas, más que una historia de jugadas o movimientos intelectuales concretos en un juego de poder. (En otro artículo, Skinner observa que incluso en el caso de que un teórico sea cínico, las razones justificatorias públicas que está obligado a ofrecer restringirán lo que puede aceptar o dejar de aceptar. Véase Quentin Skinner, «Some Problems in the Analysis of Political Thought and Action», en *Meaning and Context: Quentin Skinner and His Critics*, comp. James Tully [Princeton: Princeton Univ. Press, 1988], págs. 110-114.) Estamos, pues, de regreso al reino de los problemas intelectuales y de los intentos de resolverlos o encauzarlos.

Resulta útil tener una clasificación general de los grandes tipos de factores de que se sirven los historiadores del pensamiento para entender lo que determina y da forma a un problema. Peter Gay, *Art and Act* (Nueva York: Harper and Row, 1976), págs. 1-32, enumera tres tipos:

1. *Cultura:* factores sociales y económicos, necesidades y problemas sociales, presiones religiosas y políticas, a menudo institucionales.
2. *Artes:* las técnicas, tradiciones y herramientas de una materia o disciplina. Podemos servirnos de un término de Thomas Kuhn y llamar a eso «matriz disciplinaria»: aquellos útiles, técnicas, problemas heredados, cuerpo de conocimientos y estado presente de la discusión que son ampliamente conocidos o accesibles a los especialistas en la disciplina, así como los modelos y criterios de evaluación que hay que esperar que apliquen los participantes.
3. La *esfera privada:* la familia de la persona, la vida psicológica interior, las ansiedades, fantasías, defensas, necesidades inconscientes y *biografía* más restringidamente consideradas.

A estos factores podemos añadir otros dos:

4. Los *criterios intelectuales personales* para juzgar una teoría o para detectar un problema. (Einstein, por ejemplo, pensaba que la equivalencia de las masas gravitacional e inercial era algo que requería explicación. La existencia de una simetría allí donde no parece haber razón para esperarla, o de una asimetría allí donde parece que debería imperar la simetría; éstos y factores similares, que bordean las cuestiones estáticas, pueden plantear un problema a calibrar por un pensador.) No es necesario que esos criterios personales gocen de difusión entre los especialistas en la disciplina, pero pueden *llegar a estarlo* si el secundarlos ha llevado a desarrollar una teoría poderosa que hace que estos criterios destaquen por encima de otros.
5. *Modos generales de pensamiento* en la sociedad, no necesariamente fundados en instituciones. Eso incluye: un marco de creencias, por el estilo de

la metafísica descriptiva de Strawson; un marco de principios generales causales y explicativos; una delimitación de los tipos de cosa que necesitan explicación y de los tipos que no la necesitan; y una delimitación de los tipos de factores a los que puede apelarse como factores explicativos o como evidencia para una teoría.

Dada una determinada especificación de los componentes de una particular situación de problemas (su objetivo, sus estados y recursos iniciales, operaciones admisibles y restricciones), podemos pasar a investigar cuál de los cinco tipos de factores han dado forma a esos componentes particulares. Podemos construir una matriz de las posibilidades de influencia y, para un determinado problema, podemos investigar el modo en que cada columna ha dado forma a cada fila (por ejemplo, el modo en que la matriz disciplinaria ha fijado o dado forma a las restricciones, el modo en que la cultura ha dado forma a los objetivos, etc.). No se trata de una *teoría* del planteamiento de problemas; se trata de una categorización de los varios tipos de influencia, de una estructura, en el seno de la cual pueda organizarse la investigación histórica, de una lista-control de las cuestiones que hay que preguntarse. Podemos preguntarnos: ¿Qué es *para él* lo que hizo que *aquellos* fueran los objetivos, estados y materiales iniciales, operaciones admisibles y restricciones? ¿Y cómo consiguió dar estructura a esta situación y llegó a pensar que él mismo se estaba enfrentando a este particular problema, por muy tiznados o confusamente definidos que estén sus componentes?

La historia disciplinaria se concentra en el modo en que la matriz disciplinaria configura a la situación de problemas y, por lo tanto, a los productos intelectuales resultantes. Historias más amplias pueden abarcar los cinco factores. Mas, puesto que los fabricantes de los productos intelectuales a menudo sitúan su trabajo en relación con productos anteriores, criticándolos, o modificándolos, o desarrollándolos, dando así contenido específicamente diferencial a su propio trabajo, uno de los temas guía de la historia del pensamiento es que la matriz disciplinaria habrá de desempeñar un papel significativo. La tarea del historiador del pensamiento no termina con el estudio de la creación de una teoría o de una idea; también tiene que estudiar cómo se difunde y el impacto que tiene tanto en una disciplina, como en la sociedad toda, lo que incluye su impacto en cada uno de los cinco factores de la matriz (cultura, artes, etc.). ¿Qué contribuye a abrir espacio para una idea nueva para que pueda incluso ser considerada posible? (Véase Hans Blumenberg, *The Legitimacy of the Modern Age* [Cambridge, Mass.: M.I.T. Press, 1983], págs. 457-481.) ¿Qué es lo que determina cuánta atención se presta a la idea, quién contribuye a su propagación por la disciplina, por otras disciplinas, por la sociedad toda, y qué incentivos le mueven? ¿Quiénes ponen los micrófonos ante ciertas ideas, y por qué deciden amplificarlas? (Véase Bruno Latour, *Science in Action* [Cambridge, Mass.: Harvard Univ. Press, 1987], sobre el proceso de construcción de la red de aliados en la investigación científica.) ¿Cómo se modifica o diluye una idea a medida que se difunde? El historiador del pensamiento puede investigar también qué es lo que determina el modo en que una idea se las arregla en su competición con otras ideas en la disciplina o en la sociedad. En particular, ¿había criterios objetivos y racionales, de acuerdo con los cuales el vencedor en una competición fue superior al vencido? Aun si había criterios objetivos dentro de la disciplina que permitieran establecer que un competidor era superior a todos los demás, la amplia gama de posibles criterios significa que debemos seguir investigando por qué se apeló a estos particulares criterios.

44. Pero incluso esos cuadernos pueden haber estado sometidos a trabajo de edi-

ción y de «limpieza» por parte del mismo pensador. Tal fue el caso de Miguel Ángel y las cartas y dibujos que dejó, un *corpus* diseñado para robustecer la imagen de sí mismo como alguien que no había aprendido nada de otros e inmaculadamente exitoso en sus proyectos.

45. Véase Pat Langley, Herbert Simon, Gary Bradshaw y Jan Zytkin, *Scientific Discovery* (Cambridge, Mass.: M.I.T. Press, 1987), págs. 3-36, 49-59; D.N. Perkins y Gavriel Salomon, «Are Cognitive Skills Context-Bound?», *Educational Researcher* 18, n. 1 (enero-febrero 1989): 16-25.

46. Véase Frank Ramsey, *The Foundations of Mathematics and Other Logical Essays* (Londres: Routledge and Kegan Paul, 1931), págs. 115-116.

47. Para una discusión de la asimetría y la simetría en el pensamiento de Einstein, véase Gerald Holton, «On Trying to Understand Scientific Genius», reproducido en su *Thematic Origins of Scientific Thought* (Cambridge, Mass.: Harvard Univ. Press, 1973), págs. 353-380.

48. John Holland, Keith Holyoak, Richard Nisbett y Paul Thagard, *Induction: Processes of Inference, Learning, and Discovery* (Cambridge, Mass.: M.I.T. Press, 1986), págs. 286-319.

49. Sobre este punto, véase Howard Gardner, *The Creators of the Modern Era* (en prensa).

50. Simon y Newell, *Human Problem Solving.*

51. Véase Georg Polya, *Patterns of Plausible Inference,* 2.ª ed. (Princeton: Princeton Univ. Press, 1986).

52. Véase Kenneth Arrow, *Social Choices and Individual Values* (Nueva York: John Wiley, 1951) [trad. cast.: *Elección social y valores individuales*, Madrid, Ministerio de Economía y Hacienda, 1974]; Amartya Sen, «Social Choice Theory», en *Handbook of Mathematical Economics,* comp. K.J. Arrow y M. Intriligator (Amsterdam: North Holland, 1985); John Milnor, «Games against Nature», en *Decision Processes,* comp. R.M. Thrall, C.H. Coombs y R.L. Davis (Nueva York: John Wiley, 1954), págs. 49-60; Luce y Raiffa, *Games and Decisions,* págs. 286-298.

53. Para un ejemplo muy modesto, véase la discusión de la estructura r × H para el castigo retributivo en mis libros *Anarchy, State, and Utopia* (Nueva York: Basic Books, 1974), págs. 59-64, y *Philosophical Explanations* págs. 363-380, 388-390. Lo que importa es que incluso una estructura tan trivialmente simple puede generar resultados interesantes. La teoría de la justicia de la titulación desarrollada en *Anarchy, State, and Utopia* es otro ejemplo de modelo modesto construido por analogía con la estructura general de un sistema formal (con axiomas, reglas de inferencia y teoremas resultantes).

54. Véase Langley, Simon, Bradshaw y Zytkin, *Scientific Discovery.*

55. Véase mi «Newcomb's Problem and Two Principles of Choice», en *Essays in Honor of C.G. Hempel,* comp. N. Rescher y otros (Dordrecht: Reidel, 1969), págs. 135-136.

56. Para discusiones recientes de los experimentos intelectuales, véase Nancy Nersessian, «How Do Scientists Think?» en *Cognitive Models of Science,* comp. Ronald Giere (Minneapolis: University of Minnesota Press, 1992), esp. págs. 25-35, y David Gooding, «The Procedural Turn», en ibíd., esp. págs. 69-72.

57. Véase Thomas Kuhn, «Objectivity, Value Judgment, and Theory Choice», en Kuhn, *The Essential Tension* (Chicago: Univ. of Chicago Press, 1977) [trad. cast.: *La tensión esencial*, Madrid, FCE, 1983], págs. 331-332.

58. F. A. Hayek, *The Constitution of Liberty* (Chicago: Univ. of Chicago Press, 1960), cap. 2.

59. Sobre la limitación en nuestro grado de alerta véase mi *The Examined Life,*

págs. 40-42. Fue Hayek quien definió el grado de civilización como la medida en que nos beneficiamos del conocimiento que nosotros mismos no tenemos.

60. Véanse R. Boyd y P.J. Richerson, *Culture and the Evolutionary Process* (Chicago: Univ. of Chicago Press, 1985); John Tooby y Leda Cosmides, «Evolutionary Psychology and the Generation of Culture», *Ethology and Sociobiology* 10 (1989): 29-49; Alan Gibbard, *Wise Choices, Apt Feelings.*

61. Al aprender de otros, parece que damos por supuesto que son racionales —suficientemente racionales para que nosotros comprendamos lo que traman—. ¿Hay una base en la evolución para el principio de caridad en la traducción que ya hemos analizado críticamente? No obstante, no es necesario que dicho principio sea tan general que se pueda aplicar a todos; sería suficiente con dar por supuesta la racionalidad en el propio grupo.

62. Véase Ludwig Wittgenstein, *Philosophical Investigations* (Oxford: Basil Blackwell, 1953) [trad. cast.: *Investigaciones filosóficas*, Barcelona, Crítica, 1988]; Quine, *Word and Object* [trad. cast.: *Palabra y objeto*, Cerdanyola, Labor, 1968]; Hilary Putnam, «The Meaning of "Meaning"», en su *Mind, Language and Reality: Philosophical Papers, vol. 2* (Cambridge: Cambridge Univ. Press, 1973), págs. 215-272.

63. Adam Smith, *The Wealth of Nations* [trad. cast.: *La riqueza de las naciones*, Madrid, Alianza, 1994], libro 1, capítulo 2.

64. Max Weber, *Economy and Society* (Nueva York: Bedminster, 19687) [trad. cast.: *Economía y sociedad*, Madrid, FCE, [10]1993].

ÍNDICE ANALÍTICO

ÍNDICE DE NOMBRES

La naturaleza de la racionalidad
Robert Nozick

Paidós Básica

La racionalidad es un componente crucial de la imagen que de sí misma tiene la especie humana, no simplemente una herramienta para adquirir conocimiento y perfeccionar nuestras vidas y nuestra sociedad. Pero, hoy en día, el estudio de la racionalidad se ha convertido en un asunto técnico, a veces en cosa de matemáticos, de economistas y de filósofos, de manera que la bibliografía más relevante sobre el tema aparece últimamente sumergida en prohibitivas fórmulas salpicadas de raras notaciones simbólicas con las que se elaboran estructuras matemáticas.

Robert Nozick, en la presente obra, no lamenta en absoluto este giro, pero a la vez está convencido de que el público en general necesita también comprender las exposiciones que se realicen acerca de estos asuntos. Es evidente que ni la más clara de ellas, si quiere transmitir las ideas esenciales con cierto esmero, podrá evitar algunas descripciones y algunos desarrollos técnicos. Pero el asunto de la racionalidad no puede hurtarse a la mayoría de la gente, por lo que Nozick, en su libro, trata de minimizar esos detalles demasiado complejos o confinarlos, al menos, a secciones específicas. Porque, para la salud intelectual de nuestra sociedad, es imprescindible que las ideas fundamentales sigan siendo públicas.

Robert Nozick es profesor de Filosofía en la Universidad de Harvard. Entre sus obras pueden citarse *Philosophical Explanations*, *Meditaciones sobre la vida* y *Anarquía, Estado y utopía*, esta última ganadora del National Book Award (Premio Nacional del Libro) en 1975.


ISBN 84-493-0138-6



Diseño: Mario Eskenazi